NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

SEQUENTIAL PATTERN DETECTION AND TIME SERIES MODELS FOR PREDICTING IED ATTACKS

by

William B. Stafford

March 2009

Thesis Advisor:	Magdi Kamel
Second Reader:	Albert Barreto

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** March 2009	**3. REPORT TYPE AND DATES COVERED** Master's Thesis	
4. TITLE AND SUBTITLE Sequential Pattern Detection and Time Series Models for Predicting IED Attacks		**5. FUNDING NUMBERS**	
6. AUTHOR(S) William B. Stafford			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited		**12b. DISTRIBUTION CODE**	

13. ABSTRACT (maximum 200 words)

Improvised explosive device (IED) attacks are a significant threat to coalition forces. Defeating IEDs as weapons of strategic influence has become a major objective of Combatant Commanders and their respective Joint Task Forces. This thesis attempts to identify new approaches that can help operational forces mitigate the risk of IED attacks by identifying common sequences of events that occur before an IED attack and forecasting the number of attacks in the immediate future. Using the CARMA association rules algorithm on historical data of religious, political, and IED attack events, a model is developed to explore commonly occurring sequences of events leading to an insurgency IED attack and to predict events that are likely to occur given the sequence observed to date. Time series models are also generated to identify trends and relationships that can be helpful in forecasting future monthly IED attacks based upon previous actual historical attacks. The identified sequences and forecasts could be used to help plan troop movements, rotations, force levels, as well as allocating limited resources to address imminent threats.

14. SUBJECT TERMS Sequential Pattern Detection, Time Series, Predicting IED Attacks, Data Mining.			**15. NUMBER OF PAGES** 95
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

SEQUENTIAL PATTERN DETECTION AND TIME SERIES MODELS FOR PREDICTING IED ATTACKS

William B. Stafford
Lieutenant Commander, United States Navy
B.S., Louisiana State University, 1992

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March 2009

Author: William B. Stafford

Approved by: Dr. Magdi Kamel
Thesis Advisor

Albert Barreto
Second Reader

Dr. Dan Boger
Chairman, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Improvised explosive device (IED) attacks are a significant threat to coalition forces. Defeating IEDs as weapons of strategic influence has become a major objective of Combatant Commanders and their respective Joint Task Forces. This thesis attempts to identify new approaches that can help operational forces mitigate the risk of IED attacks by identifying common sequences of events that occur before an IED attack and forecasting the number of attacks in the immediate future. Using the CARMA association rules algorithm on historical data of religious, political, and IED attack events, a model is developed to explore commonly occurring sequences of events leading to an insurgency IED attack and to predict events that are likely to occur given the sequence observed to date. Time series models are also generated to identify trends and relationships that can be helpful in forecasting future monthly IED attacks based upon previous actual historical attacks. The identified sequences and forecasts could be used to help plan troop movements, rotations, force levels, as well as allocating limited resources to address imminent threats.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank Magdi Kamel and Buddy Barreto for their guidance and mentoring during this project.

To my wonderful wife, Tammy. Your love and support during this challenging time was invaluable; my success would not have been possible without you. I love you.

To my wonderful sons, Matthew and Jonathan. Thanks for keeping me on an even keel and always wanting to play with me, even when I had homework to do. Dad will have a lot less homework now. I love you.

THIS PAGE INTENTIONALLY LEFT BLANK

I. THESIS OVERVIEW

A. INTRODUCTION

This thesis explores data mining techniques to identify sequences of events that can be used to predict when improvised explosive device (IED) attacks will occur against coalition forces. Additionally, time series analysis examines the trends in the number of IED attacks over time and attempts to build a predictive model to forecast the number of future attacks. Further, the time series analysis attempts to identify predictor variables that can be used to improve the accuracy of the prediction model. The remainder of this chapter discusses the area of research, proposed research questions, the research methodology used, scope and limitations, and overall thesis organization.

B. AREA OF RESEARCH

An important set of techniques used in Data Mining is that of Sequence Analysis and Detection. These techniques are used to discover sequential patterns in time-oriented data. Sequence analysis and detection can be applied to any suitable data in different application domains. For example it could be used to track the performance of equipment over time, looking for patterns that lead to failure, or study credit card transactions to look for sequences that may provide hints about whether a card is renewed. This research focuses on the use of sequence analysis and detection to explore sequences of events

leading to IED attacks as well as using time series analysis techniques to forecast the number of IED attacks in the immediate future.

The objective of this research is to develop a predictive model for the timing and frequency of IED attacks. Using the CARMA association rules algorithm on historical data of religious, political, and IED attack events, a model is developed to explore commonly occurring sequences of events leading to an insurgency IED attack and to predict events that are likely to occur given the sequence observed to date. The identified sequences could be used to help plan troop movements, rotations, force levels, as well as allocating limited resources to address imminent threats. Time series models are generated to determine if there are trends or relationships that can be helpful in forecasting future monthly IED attacks based upon previous actual attacks.

Events considered for the sequence detection analysis of this effort include: Government and Infrastructure, IED, Indiscriminate, Iraqi Security Forces, Police, Religious Leaders, and Tribal Leaders attacks. The time series portion of the research establishes a baseline predictive model and then adds the predictor variables of coalition force strength and the religious holiday of Ramadan to see if those two variables influence the accuracy of the basic prediction model.

C. RESEARCH QUESTIONS

The following research questions are proposed for this research:

1. Can data mining techniques/approaches be useful in predicting the timing, frequency, and number of IED attacks?

2. Can the models be deployed to operational commands and be used by deployed personnel?

3. How accurate can these models be? Models will be evaluated for confidence, support, and goodness of fit.

D. RESEARCH METHODOLOGY

The Cross-Industry Standard Process for Data Mining (CRISP-DM), a methodology for data mining projects, was used for this research project. The methodology consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase consists of tasks and subtasks that are executed in a highly iterative and cyclical fashion. A summary of each phase follows.

1. Business Understanding

The first phase of the model is used to determine the objectives of the data mining effort, i.e., what question or problem does the analyst hope to solve with data mining.

2. Data Understanding

The data understanding phase is designed for data collection, gaining an understanding of what the data represents, determining data quality, and generating potential theories for how the data can be used to achieve the goals of the data mining project.

3. Data Preparation

The purpose of the data preparation phase is to transform the data into the format that will be imported into the modeling tool.

4. Modeling

In the modeling phase, the analyst builds, tests, and runs models that the analyst believes will provide the knowledge to answer the questions posed in the business understanding phase and meet the stated objectives of the data mining project.

5. Evaluation

The evaluation phase is used to evaluate the results of the model to see if the results meet the objectives of the data mining project. If the model did not meet the objectives of the project, then determining possible reasons why the model did not meet the objectives is an important component of the evaluation phase.

6. Deployment

This phase consists of producing a final report describing the data mining efforts as well as making decisions regarding future work. If the data mining project is successful, a repeatable data mining process for use by deployed forces in theater can be developed and implemented in this phase.

E. SCOPE AND LIMITATIONS

The scope of the thesis was constrained by the following limitations:

1. All data sources used for this research are unclassified. There is a loss of fidelity in the models due to the unclassified nature of the data. However, the goal of the thesis was to develop predictive models that are data independent so that operational users could input their own data.

2. A finite list of events and predictor variables were used in the research. The structure and quality of the available unclassified data did not allow for a more granular event classification scheme. Religious affiliation of attack victims would have been a welcome data attribute, but that level of detail was simply not available.

3. The geographic unit of analysis was limited to entire provinces. Available data did allow drilling down and conducting the analysis at a finer granular level.

F. THESIS ORGANIZATION

The remainder of this thesis is organized as follows. Chapter II gives an overview of data mining and describes the CRISP-DM methodology in more detail. Chapter III focuses on the implementation of the first three steps in the CRISP-DM methodology, namely business understanding; data understanding; and data preparation, to the problem on hand. Chapter IV discusses the remaining three steps of the CRISP-DM methodology, namely, modeling, evaluation, and deployment. Chapter V summarizes the effort of the thesis, provides conclusions, lessons learned, and makes recommendations for future work.

THIS PAGE INTENTIONALLY LEFT BLANK

II. OVERVIEW OF DATA MINING

A. INTRODUCTION

Data mining is a process that examines large data sets and attempts to find meaningful patterns and/or rules that describe how the data is related and how those patterns can be used for prediction (Berry and Linoff, 2004, pg. 7). Data mining is about knowledge discovery in large data sets. It is rapidly emerging as a valuable analysis tool in different industries and types of organizations as the volume of data that an organization produces and stores grows beyond the ability of humans to comprehend and analyze it. As an example, every time a customer uses his/her customer loyalty card at the local supermarket, the supermarket collects information on what the customer is purchasing, the date, time, day of week, etc. Since the supermarket already knows some demographic information about the customer when they filed the application to obtain the card, it can then aggregate the purchasing habits of all their customers and then use data mining tools to analyze their basket purchases. Modern computer processing power enables complex analysis of large data sets in reasonable amounts of time.

This overview chapter discusses the common types of data mining tasks in Section B, the important data mining models and algorithms to support those tasks in Section C, describes an industry-wide standardized process approach to data mining, called CRISP-DM, in Section D, and lastly, provides an overview of SPSS Clementine, the data modeling software application used for this thesis in section E.

B. DATA MINING TASKS

The majority of data mining tasks can be broken into six general areas. A brief description of those areas follows (Two Crows article; Berry and Linoff, 2004, pp. 8-12).

1. Classification

Classification is a process that seeks to group together objects with similar attributes. When classifying a new object, its attributes are compared with the attributes of objects in existing groups, and the object is assigned to the group whose attributes are most similar. An example of a classification task would be classifying customers based on their credit risk. When applying for a new consumer loan, customers are queried on multiple attributes such as salary, length of current employment, current address and how long the customer has lived in his/her current residence, what other outstanding debts does the customer have, and whether he/she rents or owns his/her current residence. Answers to these questions can be used to develop a classification model that can be applied to new loan applicants to classify their credit risk.

Data mining techniques that support classification are decision trees and nearest neighbor techniques.

2. Estimation

Estimation is used to predict the value of a target variable when this variable is continuous, such as age, salary, or weight. As an example, an automobile dealer may examine all of its auto loan applications for a certain

model car and determine an estimate of the salary range of potential buyers of that model car. Armed with this information, the auto dealer can purchase lists containing households that meet the income estimate and mail them targeted advertisements for that model of car.

Data mining techniques that are useful for estimation are regression and neural networks.

3. Association Discovery and Sequence Detection

Association Discovery is the task of determining which items in a set belong together based upon the frequency of their occurrence. Market basket analysis is classic example of association discovery. For example, when a customer purchases spaghetti sauce at the grocery store, do they also purchase some kind of meat and pasta noodles? Knowing what items a shopper tends to buy in groups can help grocery stores with targeted advertising, pricing, promotions, product location in the store, and inventory management.

Sequence detection is similar to association discovery but attempts to identify commonly occurring sequences of events over time. An antecedent is the event or sequence of events that occur prior to the consequent and are equivalent to an independent variable. The dependent variable equivalent is referred to as the consequent. Sequence pattern detection attempts to find out if the antecedent occurs, what is the probability of the consequent occurring. This sequence can be expressed as:

$$\text{Consequent} \Leftarrow \text{Antecedent}$$

There can be more than one antecedent in a given sequence and it is up to the analyst to specify the maximum number of antecedents that make up a sequence as well as other model parameters.

A widely used data mining technique for association discovery and sequence detection is the CARMA algorithm.

4. Clustering

Clustering is another method for categorizing objects based upon their attributes. The main difference between clustering and classification is that there are no preset rules about what constitutes a group of items. Clustering approaches attempt to group items together that are very similar to members of their own cluster, but vastly different from items in other clusters.

A classic example of clustering comes from the domain of astronomy. In the early 1900s, two astronomers (Enjar Hertzsprung of Denmark and Norris Russell of the United States) were trying independently to determine the relationship between a star's temperature and its luminosity (brightness). A scatter plot consisting of the temperature on one axis and the luminosity of the other axis produced a plot with three distinct clusters of stars. The characteristics of each of the three clusters were strikingly different but the individual stars in each cluster exhibited a similar relationship between temperature and luminosity to members of the same cluster. This realization led to additional research that showed the differences among the three main clusters were caused by different processes that were occurring to generate the heat and light (Berry and Linoff, 2004, pp. 351-352).

Data mining techniques that are useful are cluster detection and self-organizing maps.

5. Time Series

Time series analysis is a task for predicting future values based upon a time ordered set of previous values. A key requirement to performing time series analysis is that historical data must be measured at constant time intervals. An example of time series analysis comes from the stock market. Stocks traded on the New York Stock Exchange have well defined prices at the close of each trading day. An analyst could use time series analysis to attempt to predict future stock prices based upon historical prices.

Data mining techniques used for time series are exponential smoothing and Autoregressive Integrated Moving Average (ARIMA).

C. DATA MINING MODELS AND ALGORITHMS

There are a multitude of data mining models and algorithms contained in commercially available software applications. A description of several data mining techniques follows (Two Crows Manual, 1999; Berry and Linoff, 2004).

1. Regression

Regression models examine the relationship between sets of data to determine if one of the sets is dependent on the other set. When examining pairs of (x,y) data, a common regression technique, called linear regression, attempts to fit a straight line that best approximates the

distribution of the (x,y) coordinates. The least squares method is used to calculate the slope of the line and the y-intercept.

An example would be to determine if there is a linear relationship between the outside air temperature and the sales of ice cream at the grocery store. Both numbers could be recorded each day over the course a year and then plotted on a graph. Using the least squares method, a formula can be calculated that best fits a line with the data points. The fitted line can be used to forecast the sales of ice cream based on the outside air temperature.

Since there are very few real life problems with only one independent variable, more complex forms of regression analysis have been developed to deal with problems that have multiple independent variables. Another type of regression analysis, known as non-linear regression, is used when the dependent variable changes at a non-constant rate such as in exponential functions or logarithmic functions.

2. Decision Trees

A decision tree represents a set of hierarchical rules that allow the classification of a large group of dissimilar items or events into smaller groups that are increasingly similar the farther down the tree the item is located. An example of a decision tree is a tree that classifies new loan applicants as high risk or low risk. A simple decision tree adapted from the Two Crows article is included as Figure 1.

Figure 1. Simple Classification Tree
(From: Two Crows, p. 15).

The first rule to evaluate a credit application is the income of the prospective applicant. The rule checks whether the applicant has an income of $40,000 or greater. If the applicant has indeed an income greater than $40,000 a second rule is applied that checks his/her credit. As rule decisions are made, the loan officer travels down tree branches until reaching a leaf node where the applicant is classified as either high or low risk.

Decision trees can be created in two ways. The first way is to carefully examine the members of a heterogeneous group and determine if there are any characteristics that are common to some of the members but not to other members. For an example, in an animal kingdom classification, one could separate animals that can breathe water (fish) from those that do not (people). Another way to create decision trees is to use training data that is already classified. In the loan applicant example, the node that asks if the applicant has been in their current job for five years or

more is an example of a rule that could be created by analyzing historical records of previous loan applicants and loan defaults.

Two advantages of decision trees are that they are easy for humans to comprehend the sequence of decisions that led to the final decision and decision trees work well when the number of decisions that have to be made are of reasonably small size. Two disadvantages of decisions tree are that they do not handle ambiguity well and that if the number of decisions in the tree gets too big, the ability of the decision tree to correctly determine the classification decreases.

3. Neural Networks

Neural network models in computers have been developed to attempt mimic the decision-making and learning characteristics of biological neural networks.

A neural network consists of three layers: an input layer, a hidden layer, and an output layer. Inside the input layer are input nodes that represent the independent variables for the problem domain. The hidden layer contains hidden nodes which take the input from the input nodes and passes it to the output nodes via a weighted function. Each hidden node is connected to all of the input nodes and all of the output nodes. The output layer contains the output nodes that represent the dependent variables of the solution. A simple neural network is presented in Figure 2.

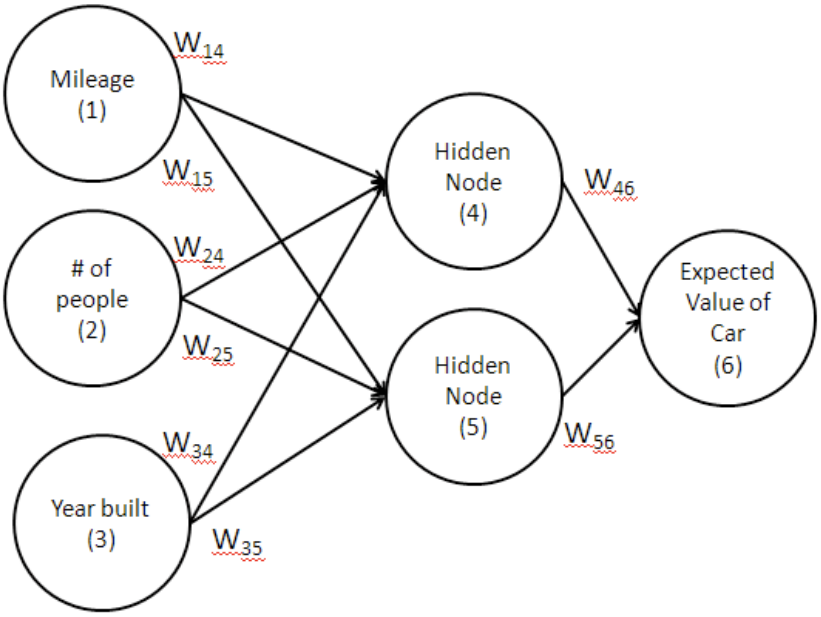


Figure 2. Simple Neural Network.

Consider the neural network of Figure 2 that is used to determine the price of a used car. The input nodes consist of three attributes that a used car possesses, such as the mileage, the number of passengers that can fit in the car, and the year the car was manufactured. The number of hidden nodes are arbitrarily set by the data analyst or by the software. For this example, two nodes are specified. The output node takes the weighted input from the two hidden nodes, applies a function to the input variables and then produces a value that represents the expected value of the used vehicle. The value that each node passes to the next node is a function of the weighted sum of the products of inputs to that node. The value of node 1 can be expressed as N_1 and the weight of the link between nodes 1 and 4 can be annotated as W_{14}. For example, the sum of the inputs to node 4 are $(N_1 \cdot W_{14}) + (N_2 \cdot W_{24}) + (N_3 \cdot W_{34})$.

Node 4 would then use that input as an input variable for its function and pass the new value to the next node (node 6 - the output node). In a similar manner, the expected value of the car (the output of node 6), can be expressed as:

$$Value = f(N_4 \cdot W_{46} + N_5 \cdot W_{56}) .$$

The initial values of the nodes and link weights can be set by the user or by the software.

The advantage of a neural network is that it can be trained using actual values, thereby "learning" how to calculate more accurate values in the future. This is done by telling the model how much error there is between its calculation and the actual value of a used car. Error in this case is defined as the difference between the estimated and actual prices. The error at each node can be calculated when the error at node 6 is known. The cumulative error at node 6 can be divided into component errors that came from each node and link. A correction factor can be applied to each node and link based upon the error. As more and more vehicle data is used in training the model, the amount of error that is observed will decrease.

A disadvantage of neural networks is that the network function is similar to a "black box" because all the node and link weighs are hidden from the analyst. There is no way to deduce how the neural network reached a decision.

4. K-Nearest Neighbor and K-means Algorithm

When humans encounter a new problem, they often attempt to solve the problem based upon similar problems

they have experienced in the past; i.e., they try solutions that have worked in the past on the new problem. This natural human tendency is the underlying theory behind k-nearest neighbor.

An example of k-nearest neighbor is detecting income tax fraud. When a new return is received, the tax return can be compared against hundreds of thousands of other returns. If the new return is closest in similarity to fraudulent returns, then the new return can be flagged for further investigation. The key principle is in determining how close the nearest neighbors are. The distance between numbers can easily be calculated, but distances between categorical values (for example married and single) take more thoughtful analysis for their determination.

An advantage of k-nearest neighbor algorithm is that it is relatively easy for humans to understand. A disadvantage of k-nearest neighbor is that it is computationally intensive as more items are added since the distance has to be calculated between the newest member and all of the previous members.

Another type of algorithm that relies on the distance between objects is the K-means algorithm, published in 1967 by J.B. McQueen. It operates as follows. Step 1 of the algorithm selects the first k-records and assign them as seeds (they are the centers of the clusters). Step 2 compares each subsequent record with each of the seeds and assigns it to the cluster that it is closest to in similarity. One possible way to determine which seed is

closer is to measure the geometric distance between the record and each seed in the same manner as k-nearest neighbor algorithm.

After all of the records have been placed in initial clusters, the geometric center can be calculated for each cluster. The geometric center then becomes the new center of the cluster (the center shifts from the initial seed position to the new geometric center of the cluster). The algorithm then starts again at step 2 and reassigns each record to the closest cluster. A new geometric center is calculated and the process repeats until the geometric centers stop shifting around. Once the geometric centers are stable, the clustering process in complete.

An advantage of the K-means algorithm is that the objects do not require any initial classification rules, the only requirement is that suitable metrics are available that allow the distance to be measured between objects.

A disadvantage of K-means algorithm clustering is that knowing what cluster an object belongs too does not usually answer the question that the data mining analyst was looking for. Typically, additional mining procedures are performed to produce useful knowledge.

D. CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)

The CRISP-DM methodology (shown in Figure 3) is a standard framework developed by a consortium of companies to support data mining projects. It consists of six major phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. A description of each phase follows (CRISP-DM Manual, 2000).

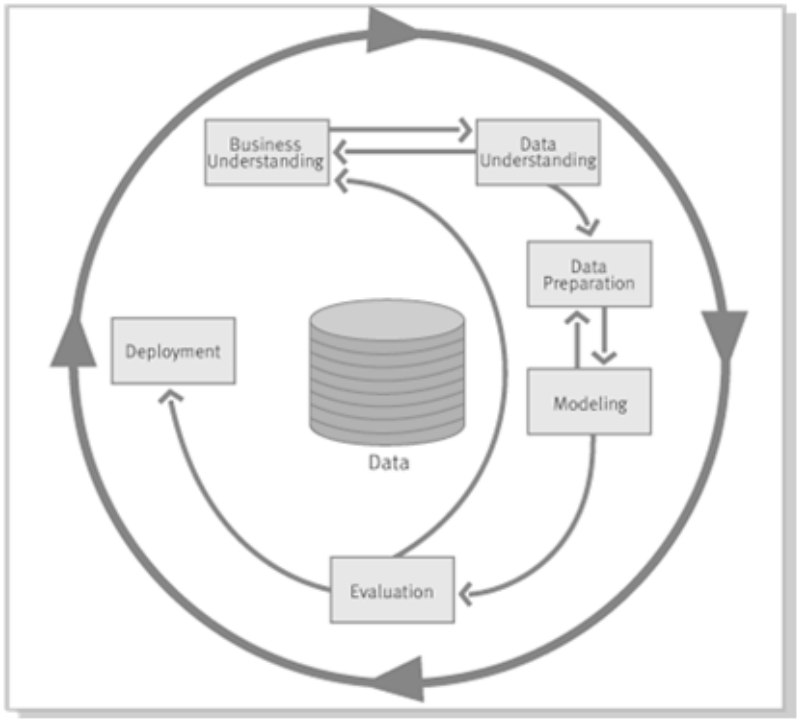


Figure 3. CRISP-DM Reference Model (From: CRISP-DM Manual).

1. Business Understanding

The first phase of the model is used to determine the objectives of the data mining effort, i.e., what question or problem does the analyst hope to solve with data mining.

The first sub-task in this step is to determine the business objectives. An example objective might be to increase the sales of a certain model of vehicle. The subsequent sub-task is to determine the data mining data-mining task that will accomplish the business objective. For example, determining what type of customers bought that model vehicle in the past.

The next step is to assess the current situation by answering the following questions. What types of assets

are available to use in the project, what are the constraints of the project - such as funding or time. Are there any risks or consequences associated with the project? A cost/benefit analysis can be performed to verify that the return on investment for the data mining project justifies the cost that will be incurred.

The last sub-task is to determine an initial project plan that will support discovering the knowledge that will answer the business question.

2. Data Understanding

The data understanding phase is designed for data collection, gaining an understanding of what the data represents, determining data quality, and generating potential theories for how the data can be used to achieve the goals of the data mining project.

The first step in the data understanding phase is to collect the initial data that will be analyzed during the modeling phase.

The next step is to describe the data and gain an understanding of what the data represents. This is followed by an initial exploration of the data to determine if there are any obvious patterns or relationships in the data. Performing basic statistical analysis of the data can lead to initial hypotheses for further analysis.

The last step in the data understanding phase is to examine the data to determine if there are any data quality issues such as data that is incorrect, missing, incomplete, or ambiguous.

For the effort of this thesis, the initial data collection efforts focused on coalition force casualties, Iraqi casualties, IED events, and religious holidays. All data was collected from unclassified sources. A detailed description of the data collection effort and associated data sources is presented in Chapter III.

3. Data Preparation

The purpose of the data preparation phase is to transform the data into the format that will be imported into the modeling tool. The first step of this phase is to decide which data will be used, since not all of the data may be of suitable quality for inclusion, or that a subset of the data is sufficient for the model.

The next step is to clean the data, which involves using methods to increase the quality of the data such as predicting missing values or adding suitable defaults for missing values. This is followed by data construction which involves creating new records from existing data, modifying values of attributes, or creating derived values from existing values.

After the data construction step is complete, data is integrated. Data integration is used to create new records from multiple tables (an example is combining sales data with demographic data to produce a table that showed sales behavior by customer demographics).

Lastly, the data needs to be formatted so that it will be accepted by the modeling tool. Format changes do not change the meaning of the data, just the way the data is presented. An example might be changing dates from a

month-day-year format commonly used in the United States to the day-month-year format that is used in European countries.

Data preparation for this project was considerable. Data was imported into Excel for initial cleaning and formatting. It was then imported into the data mining tool for further manipulation and preparation. A detailed description of the data preparation is presented in Chapter III.

4. Modeling

In the modeling phase, the analyst builds, tests, and runs models that the analyst believes will provide the knowledge to answer the questions posed in the business understanding phase and meet the stated objectives of the data mining project.

The phase starts with model selection that the analyst deems most appropriate for the objectives that have been set forth for the data mining project. The analyst then builds some test models and determines what results the test models produce with test or sample data. If the test models produce the expected results, then the analyst builds the full model for the data mining project. Building the model involves determining model parameters and optimizing the model for performance or results. The last step is to assess the model to determine how the model performed.

The modeling tool selected for this project was SPSS Clementine and the specific models are sequence pattern detection and time series analysis.

5. Evaluation

The evaluation phase is used to evaluate the results of the model to see if the results meet the objectives of the data mining project. If the model did not meet the objectives of the project, then determining possible reasons why the model did not meet the objectives is an important component of the evaluation phase.

The evaluation phase consists of three steps, the first of which is to evaluate the results of the model. Do the results produce answers to the questions posed in previous project phases? After evaluating the results, the model creation process is studied to determine if the analyst inadvertently left out a critical factor or process that would improve the model if included. Lastly, decisions need to be made on what actions are next for the project. If the model is good and producing valuable knowledge, then proceeding to the deployment phase is often a good idea. If the model is not producing any new knowledge, then the model can be modified to try and improve it or perhaps even abandoned to try again from scratch with fresh perspective.

The sequence pattern detection models did find some sequences with reasonable support and confidence, but domain experts in the field would have to decide whether the data was useful.

6. Deployment

Deployment is the last phase of the CRISP-DM methodology and the phase where the data or knowledge produced by the data mining project is converted into a

format that is useable by the operator in the field. The output may be as simple as a written report detailing the results of the project to a self-contained computer application that can create subsequent models based upon new data submitted by the operator in the field.

It is important that the results of the project (whether it is the written report or the computer application) are fully understood by the operator in the field since the analyst will not be available at every location to explain what the report means or how the application works.

The efforts of this thesis focus on the first four steps of the CRISP-DM methodology. Ideas on the evaluation and deployment of the generated models are presented in Chapter IV. Recommendations for future work including deployment are included in a later chapter.

E. SPSS CLEMENTINE OVERVIEW

The data mining tool used for this thesis is SPSS Clementine. Clementine is a data mining workbench that enables the development of predictive models and their deployment into the operational environment of organizations to improve decision making. Designed around the industry-standard CRISP-DM model, used in this research, Clementine supports the entire data mining process, from data input to actionable results.

Data mining using Clementine is based on the process of running data through a series of nodes, referred to as a stream. This series of nodes represents operations to be performed on the data, while links between the nodes

indicate the direction of data flow. Typically, a data stream is used to read data into Clementine, run through a series of manipulations, and then sent to a destination, such as a file or report. A screen capture of a simple representative stream is presented as Figure 4.

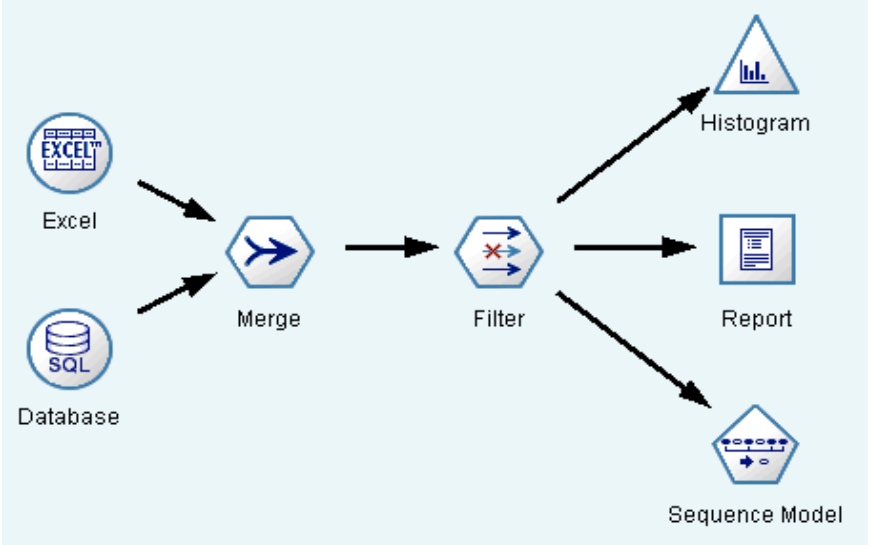


Figure 4. Sample SPSS Clementine Stream.

The two nodes of the left of Figure 4 are source nodes. Excel and relational databases are two potential sources, others include delimited column text files, fixed width text files, and user input. The merge node is a type of record operations node and it allows merging of records from multiple inputs. Aggregate, sort, append, and distinct are some of the other record operations nodes that are available.

The filter node in the middle of the figure is an example of a field operations node that allows fields to be

renamed or removed. Some of the other types of field operations nodes are derive, type, reclassify, binning, and partition.

The histogram node is a graph node that displays a frequency graph of values. The other types of graphs available include plot, distribution, collection, and time plot.

The report node is an output node that creates a report from the data. Other report node options are table, matrix, data audit, and statistics.

The sequence model node is a modeling node that generates sequence pattern detection models that are discussed in this thesis. Neural net, time series, K-means, feature selection, and regression are some of the other available modeling choices.

The next chapter addresses the implementation of the first three steps of the methodology to the IED problem by describing the business drivers and objectives of the IED problem, the process of collecting and understanding the data required for data mining, and the preparation and manipulation of data for the modeling phase of the methodology.

III. BUSINESS UNDERSTANDING, DATA UNDERSTANDING, AND DATA PREPARATION

A. INTRODUCTION

This chapter addresses the implementation of the first, second, and third steps of the CRISP-DM methodology, namely business understanding; data understanding; and data preparation, to the IED problem.

Specifically, the chapter is organized as follows. Section B addresses research objectives and requirements from a business (operational) perspective, and converts this knowledge into a data mining problem definition. Section C describes initial data collection with particular emphasis on the content and format of data sources used for this research. Finally, Section D discusses all the activities undertaken to construct the final dataset (data that will be fed into the modeling tool) from the initial raw data.

B. BUSINESS UNDERSTANDING

After the war in Iraq ended, United States and Coalition partners transitioned to security, stability, and reconstruction operations to assist the Iraqi government with establishing a self-sustaining government. An insurgency developed to stop the Coalition Forces from accomplishing their objectives.

As the insurgents do not have access to high tech weapons, they use lower technology weapons that are inexpensive to manufacture and do not require extensive training to deploy. An improvised explosive device (IED)

is a relatively simple device to build but can cause great harm when employed. An IED is essentially a "home-made" bomb that can be built using explosives and a detonator. IEDs are often disguised as innocent looking objects or are hidden out of sight to increase the likelihood of a Coalition member being hurt or killed when the device is discovered.

To counter the use of IEDs in Iraq by insurgents, the United States Army established the Army IED Task Force in October 2003, which has evolved into the Joint Improvised Explosive Device Defeat Organization (JIEDDO). JIEDDO has a three-pronged approach to countering IEDs:

- Attack the Network
- Defeat the Device
- Train the Force

Toward this end, JIEDDO has organized the requirement needs by five different functional areas: (JIEDDO, 2009)

- Predict/Prevent
- Detect
- Neutralize
- Mitigate
- Training

JIEDDO has partnered with commercial entities for manufacture of equipment to support their goals as well as with academic institutions for research.

Countering the IED threat is the operational objective of JIEDDO and this objective can be expressed as reducing or even eliminating the number of casualties that occur from IED attacks. From the CRISP-DM perspective the objective can be restated as reducing the number of

casualties caused by IEDs. Using the data mining technique of sequence pattern detection, it may be possible to determine sequences of events that precede IED attacks. Time series analysis can also be used to forecast the number of IED attacks in the immediate future based on historical actual data as well as predictor variables and intervention effects.

C. DATA UNDERSTANDING

Since this research was conducted in the unclassified domain, only unclassified publicly available data sources were used. The following data sources were used for this research:

1. Coalition IED Events

Coalition IED events are generated from http://icasualties.org/Iraq/index.aspx. The site contains links to the United States Department of Defense, United Nations, and other coalition partners news releases of the actual casualty report for the event.

Statistics for U.S. force casualties are broken down by hometown of individual (both state and city lists are available), by service, by ethnicity, and by what base the service member was originally from. Statistics are also available for coalition deaths by nationality, whether or not the fatality was in a hostile or non-hostile environment, deaths caused by improvised explosive devices, and which province in Iraq the fatality occurred in.

Additionally, military deaths are broken down by time periods - the time periods are defined by the icasualties.org staff, as well as the traditional

year/month by month, and an aggregated table showing the totals for each year. Statistics regarding U.S. forces that have been wounded is displayed showing whether medical air transportation was required along with a monthly and weekly breakdown of when the injuries where sustained. Current news reports from a variety of sources that pertain to Iraq are displayed. A screen capture of the main page of the Web site taken in February 2009 is presented as Figure 5.

Figure 5. Icasualties.org Home Page Screen Capture.

2. Iraqi Casualties Events

Iraqi casualty events (Government, Indiscriminate, Iraqi Security Force, Police, Religious Leaders, Tribal Leaders) are generated from http://www.iraqbodycount.org/. The website includes data of Iraqi incidents and causalities and is "drawn from cross-checked media reports,

hospital, morgue, NGO and official figures to produce a credible record of known deaths and incidents" (iraqbodycount.org, 2009).

The website contains data about past incidents by date along with a description of the incident as well as data about the individuals that were involved in the events. The data for incidents and individuals is cross-referenced, providing a clearer picture for each event. The site also contains a number of analysis articles written by staff members along with graphs of the number of casualties that occurred in each year since 2003. A screen capture of the main page of the Web site taken in February 2009 is presented as Figure 6.

Figure 6. Iraqbodycount.org Home Page Screen Capture.

3. Economic, Public Opinion, and Security Data

General, economic, public opinion, and security data is obtained from the Saban Center for Middle East Policy, Iraqi Index located at http://www.brookings.edu/saban/iraq-index.aspx. This site contains "information on various criteria, including crime, telephone and water service, troop fatalities, unemployment, Iraqi security forces, oil production, and coalition troop strength." The site is updated monthly and archived reports are available on the website beginning in November 2003. A screen capture of the main page of the Web site taken in February 2009 is presented as Figure 7.

Figure 7. Saban Center Iraq Index Home Page Screen Capture.

4. Geographic and Location Data

Geographic and location data are obtained from a variety of publicly available websites, such as http://www.fallingrain.com/world/IZ/. The fallingrain website contains a listing of all provinces in Iraq and the city names are organized in a hierarchy by the first letter of the city name. Selecting the first letter or symbol of a location name, narrows down the choices in the location name until reaching a web page that has an alphabetical list of locations that started with the previously selected letters. A screen capture of the page that displays Iraq locations that start with "Chi" taken in February 2009 is presented as Figure 8.

Places in Iraq that start with Chi

Up

Name	What	Region	Country	Lat	Long	Elev Ft.	Pop Est
Chia Raza	city	Muhafazat Diyala	Iraq	34.7811111	45.67	1614	7118
Chia Rida	city	Muhafazat Diyala	Iraq	34.7811111	45.67	1614	7118
Chia Surkh	city	Muhafazat Diyala	Iraq	34.6747222	45.5891667	1184	5576
Chichah Qal`ah	city	Muhafazat as Sulaymaniyah	Iraq	35.11	45.5216667	2828	9794
Chichan	city	Muhafazat Diyala	Iraq	33.9	44.9333333	173	5629
Chichān	city	Muhafazat Diyala	Iraq	33.9	44.9333333	173	5629
Chigan	city	Muhafazat Ninawa	Iraq	36.65	42.9	761	5293
Chigha Mira	city	Muhafazat Arbil	Iraq	35.7816667	43.7763889	944	8789
Chighah Mirah	city	Muhafazat Arbil	Iraq	35.7816667	43.7763889	944	8789
Chighalok	city	Muhafazat Arbil	Iraq	35.8016667	43.7219444	994	8785
Chigān	city	Muhafazat Ninawa	Iraq	36.65	42.9	761	5293
Chil Heweza	city	Muhafazat Arbil	Iraq	35.9025	43.7816667	1210	8775
Chil Hewēza	city	Muhafazat Arbil	Iraq	35.9025	43.7816667	1210	8775
Chil Juwayzah	city	Muhafazat Arbil	Iraq	35.9025	43.7816667	1210	8775
Chilbasar	city	Muhafazat Arbil	Iraq	35.9186111	44.2827778	1318	8771
Chilparat	city	Muhafazat Ninawa	Iraq	36.8269444	42.1405556	1269	5310
Chilpārāt	city	Muhafazat Ninawa	Iraq	36.8269444	42.1405556	1269	5310

Figure 8. Fallingrain Screen Capture.

5. Iraqi Province Data

Detailed information on Iraqi provinces is taken from the Multi-National Force-Iraq web page http://www.mnf-iraq.com/index.php?option=com_content&task=view&id=1469&Itemid=78. A screen capture taken in February 2009 is presented as Figure 9.

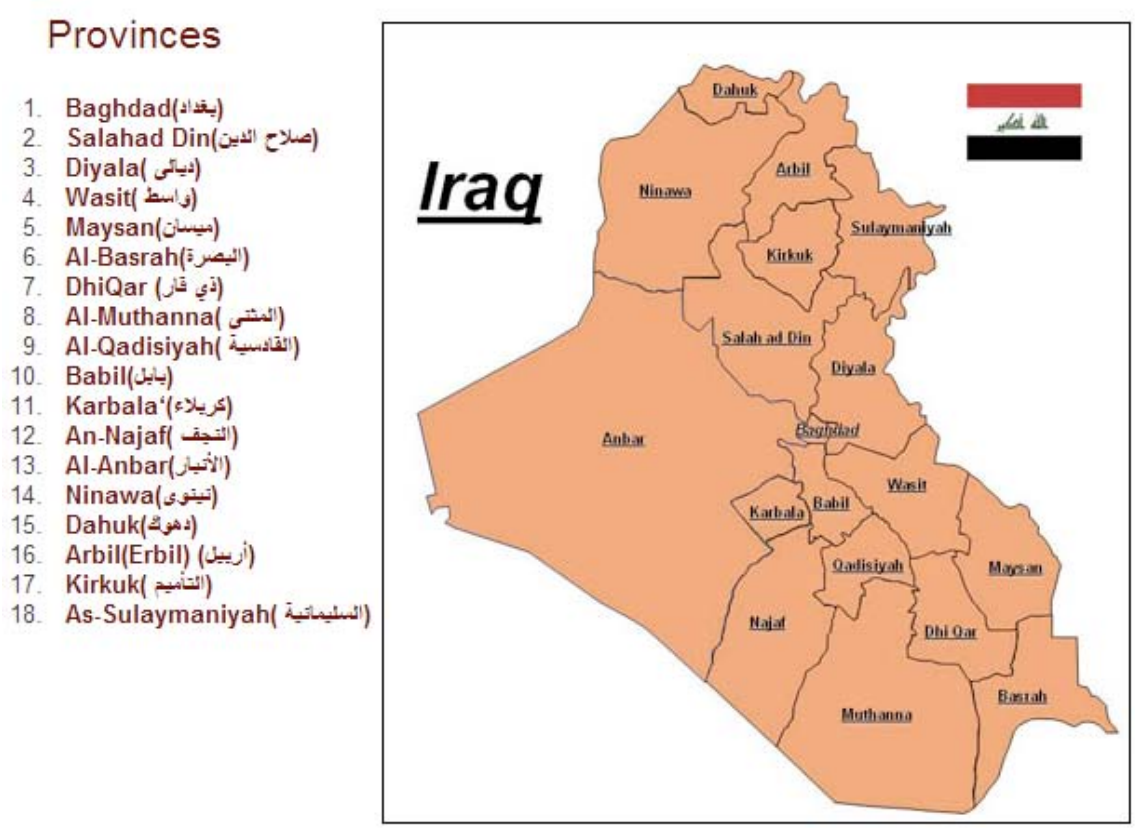


Figure 9. Provinces of Iraq.

6. Religious Holiday Data

Religious holiday data are obtained from a variety of sites, such as http://www.religionfacts.com/islam/holidays.htm.

7. Force Level Data

Coalition force level data was obtained from the Iraq Index report, published by the Saban Center for Middle East

Policy. The index authors cite numerous news articles over the last several years as the source of the data.

D. DATA PREPARATION

1. File Format for Data Mining

Although data preparation is a single step in the CRISP methodology and therefore seem to comprise about 20% of the total effort (1 of 6 phases), in practice preparing data is the most time-consuming phase of any data mining project. This data mining effort was no exception with an estimated sixty to seventy percent of the total effort spent on preparing data for mining. It is important to emphasize that clean data that is well understood is essential to the success of any data mining effort.

The goal of the data preparation phase is to merge, transform, and manipulate the data from the different sources into a format suitable for mining for the task on hand.

a. Pattern Sequence Detection

For the sequence discovery mining process, the format required by the modeling algorithm consists of a file with three fields as described in the sections below.

(1) ID field. Each unique value of this field indicates a specific unit of analysis. In this effort the unit of analysis is the geographic region or province. The official eighteen provinces of Iraq were used as values for this field.

(2) Time field. The time field indicates the date that an event has occurred.

(3) Target Field. This field contains the events of interest in sequence modeling. In this effort they represented one of eight events of interest (IED event, Iraqi Security Force event, Indiscriminate event, etc.)

b. Time Series

For the time series analysis, the format required by the modeling algorithm consists of a data file with a minimum of two fields, a time field and target (dependent variable) field. Additional predictor variables, events, and intervention fields may be included to potentially improve the forecasting accuracy of the time series model. For this research, up to four fields were used: a time field, a target field, an event field, and a predictor field.

(1) Time Field. The time field indicates the date that an event has occurred. The data used in the model must have been measured at regular intervals without any missing values. Time series analysis algorithms do not interpolate for missing values. The range of dates used were from September 2003 to November 2008.

(2) Predictor Field. Predictors are series of numeric values that may help explain the dependent series. Coalition force levels was used in some models as a predictor variable for the number of IED attacks per month.

(3) Event Field. An event is a recurring event that happens with a predictable frequency. It is represented as a boolean field. The occurrence of the Islamic holy month of Ramadan was used as an event in some of the time series models.

(4) Intervention Field. Interventions are special one-time events that do not occur with any predictable regular frequency. An example of an intervention would be the flooding of a major city due to a dam failing. No intervention events were used in this research.

(5) Dependent Field. The dependent field is the field that the model is trying to determine the pattern for and to predict what near term values for the field might be. The dependent variable is this research is the number of IED attacks per month.

2. Data Preparation/Manipulation

To achieve the format required by the data mining model, extensive manipulation on the data sources was undertaken. The following is a summary of the main data preparation effort for each model.

a. Sequence Pattern Detection

Two main sources of data were used for the sequence detection model. Data on coalition forces was obtained from icasualties.org and the data regarding Iraqi deaths was obtained from iraqbodycount.org.

(1) IED Casualty Data. Data obtained from icasualties.org consisted of deaths caused by IED and is presented in tabular format with columns of included Date, Name, Place of Death - Province, and Cause of Death. The individual names are hyperlinked to the U.S. Department of Defense Public Affairs website that includes a press

release describing the details of the attack. A partial screen capture of data for the month of April 2008 is presented as Figure 10.

Date	Name	Place of Death - Province	Cause of Death
30-Apr-2008	Specialist Ronald J. Tucker	Baghdad (southern part)	Hostile - hostile fire - IED attack
30-Apr-2008	Captain Andrew. R. Pearson	Baghdad (southern part)	Hostile - hostile fire - IED attack
30-Apr-2008	Sergeant 1st Class Lawrence D. Ezell	Baghdad (northern part)	Hostile - hostile fire - IED attack
30-Apr-2008	Staff Sergeant Chad A. Caldwell	Mosul - Ninawa	Hostile - hostile fire - IED attack
29-Apr-2008	Staff Sergeant Bryan E. Bolander	Baghdad	Hostile - hostile fire - IED attack
24-Apr-2008	Staff Sergeant Shaun J. Whitehead	Iskandariyah - Babil	Hostile - hostile fire - IED, small arms fire
22-Apr-2008	Lance Corporal Jordan C. Haerter	Ramadi (near) - Al Anbar Province	Hostile - hostile fire - IED attack (VBIED)
22-Apr-2008	Corporal Jonathan T. Yale	Ramadi (near) - Al Anbar Province	Hostile - hostile fire - IED attack (VBIED)
21-Apr-2008	Specialist Steven J. Christofferson	Baiji - Salah Ad Din	Hostile - hostile fire - IED attack
21-Apr-2008	Sergeant Adam J. Kohlhaas	Baiji - Salah Ad Din	Hostile - hostile fire - IED attack
21-Apr-2008	1st Lieutenant Matthew R. Vandergrift	Basra - Basrah	Hostile - hostile fire - IED attack
18-Apr-2008	Specialist Benjamin K. Brosh	Balad - Salah Ad Din	Hostile - hostile fire - IED attack (VBIED)
18-Apr-2008	Specialist Lance O. Eakes	Baghdad (north of)	Hostile - hostile fire - IED attack
14-Apr-2008	Sergeant Joseph A. Richard III	Baghdad (northeastern part)	Hostile - hostile fire - IED attack
14-Apr-2008	Specialist Arturo Huerta-Cruz	Tuz - Salah Ad Din	Hostile - hostile fire - IED attack
14-Apr-2008	Corporal Richard J. Nelson	Al Anbar Province	Hostile - hostile fire - IED attack
14-Apr-2008	Lance Corporal Dean D. Opicka	Al Anbar Province	Hostile - hostile fire - IED attack
12-Apr-2008	Specialist William E. Allmon	Baghdad	Hostile - hostile fire - IED attack
09-Apr-2008	Technical Sergeant Anthony L. Capra	Baghdad	Hostile - hostile fire - IED attack
09-Apr-2008	Sergeant Jesse A. Ault	Tunis (died in Baghdad) - Salah Ad Din	Hostile - hostile fire - IED attack
09-Apr-2008	Sergeant Shaun P. Tousha	Baghdad	Hostile - hostile fire - IED attack

Figure 10. Partial IED Attacks Screen Capture.

The data represented by Figure 10 was imported into an Excel table for the initial data preparation efforts.

The date field is required and sufficient for the model and therefore does not need further manipulation. The "Place of Death" field, however, which corresponds to the ID field in our model, is not in the

correct format since the location text may consist of more information that just the province. In some cases, the province is not explicitly stated at all, though it might be inferred. A new field is created that contains the corresponding province so as a single region is assigned to the event.

The location/province data from the icasualties.org site is straightforward and it is usually relatively easy to decide which province the event occurred in. For example, if the Place of Death is listed as "Baghdad (southern part)" the Region Assigned would be "Baghdad". The location values across the data set varied enough to prevent automating the process and manually entering the Region Assigned data ensured standardization of the values in this field across the dataset. Since this data included only IED attacks events, the value "IED Attack" is assigned to the content field of all the records.

A screen capture of the Excel table that was created from the icasualties.org data presented in Figure 10 is included as Table 1.

Date	Name	Place of Death - Province	Region Assigned	Cause of Death	Event Assigned
30-Apr-08	Specialist Ronald J. Tucker	Baghdad (southern part)	Baghdad	Hostile - hostile fire - IED attack	IED Attack
30-Apr-08	Captain Andrew. R. Pearson	Baghdad (southern part)	Baghdad	Hostile - hostile fire - IED attack	IED Attack
30-Apr-08	Sergeant 1st Class Lawrence D. Ezell	Baghdad (northern part)	Baghdad	Hostile - hostile fire - IED attack	IED Attack
30-Apr-08	Staff Sergeant Chad A. Caldwell	Mosul - Ninawa	Ninawa	Hostile - hostile fire - IED attack	IED Attack
29-Apr-08	Staff Sergeant Bryan E. Bolander	Baghdad	Baghdad	Hostile - hostile fire - IED attack	IED Attack
24-Apr-08	Staff Sergeant Shaun J. Whitehead	Iskandariyah - Babil	Babil	Hostile - hostile fire - IED, small arms fire	IED Attack
22-Apr-08	Lance Corporal Jordan C. Haerter	Ramadi (near) - Al Anbar Province	Al-Anbar	Hostile - hostile fire - IED attack (VBIED)	IED Attack
22-Apr-08	Corporal Jonathan T. Yale	Ramadi (near) - Al Anbar Province	Al-Anbar	Hostile - hostile fire - IED attack (VBIED)	IED Attack
21-Apr-08	1st Lieutenant Matthew R. Vandergrift	Basra - Basrah	Al-Basrah	Hostile - hostile fire - IED attack	IED Attack
21-Apr-08	Sergeant Adam J. Kohlhaas	Baiji - Salah Ad Din	Salahad Din	Hostile - hostile fire - IED attack	IED Attack
21-Apr-08	Specialist Steven J. Christofferson	Baiji - Salah Ad Din	Salahad Din	Hostile - hostile fire - IED attack	IED Attack
18-Apr-08	Specialist Benjamin K. Brosh	Balad - Salah Ad Din	Salahad Din	Hostile - hostile fire - IED attack (VBIED)	IED Attack
18-Apr-08	Specialist Lance O. Eakes	Baghdad (north of)	Baghdad	Hostile - hostile fire - IED attack	IED Attack
14-Apr-08	Specialist Arturo Huerta-Cruz	Tuz - Salah Ad Din	Salahad Din	Hostile - hostile fire - IED attack	IED Attack
14-Apr-08	Sergeant Joseph A. Richard III	Baghdad (northeastern part)	Baghdad	Hostile - hostile fire - IED attack	IED Attack
14-Apr-08	Corporal Richard J. Nelson	Al Anbar Province	Al-Anbar	Hostile - hostile fire - IED attack	IED Attack
14-Apr-08	Lance Corporal Dean D. Opicka	Al Anbar Province	Al-Anbar	Hostile - hostile fire - IED attack	IED Attack
12-Apr-08	Specialist William E. Allmon	Baghdad	Baghdad	Hostile - hostile fire - IED attack	IED Attack
9-Apr-08	Technical Sergeant Anthony L. Capra	Baghdad	Baghdad	Hostile - hostile fire - IED attack	IED Attack
9-Apr-08	Sergeant Shaun P. Tousha	Baghdad	Baghdad	Hostile - hostile fire - IED attack	IED Attack
9-Apr-08	Sergeant Jesse A. Ault	Tunis (died in Baghdad) - Salah Ad Din	Salahad Din	Hostile - hostile fire - IED attack	IED Attack

Table 1. April IED Events (Partial listing).

(2) Iraqi Casualty Data. Data regarding attacks against Iraqi government/political/awakening council members, Iraqi Security Forces, Iraqi Police Forces, Iraqi Religious Leaders, and Iraqi Tribal Leaders were obtained from http://www.iraqbodycount.org. The Incidents table is the primary table of interest for this effort. It contains the following ten fields: IBC code, Start Date, End Date, Time, Location, Target, Weapons, Reported Minimum, Reported Maximum, and Sources.

As indicated, the fields of interest for the sequence pattern detection model are the Start Date, Location (ID field), and Target. The Start Date field represents the date that the event started, and a majority of the events in the table started and ended on the same day. If the start date and end date for the event are different, the start date is used as the date for the event modeling purposes. The location field generally contains more detailed location information (for example, city names). The location information requires parsing to determine which province the event occurred in. As with the IED Casualty data, several resources are consulted to determine what provinces the events occurred in when the content of that field are ambiguous, and the regions (provinces) that are assigned are the same as the region set from the IED casualty data.

The Target field is used to determine the type of event that occurred for the purposes of the sequence pattern detection. An initial analysis of the data showed that the different types of attack targets could be mapped to the following categories: IED Attack,

Government, Indiscriminate, Iraqi Security Forces, Police, Religious Leader, Tribal Leader, and Unknown attacks. A list of the categories used along with a description of the categories is included as Table 2.

Categories	Description
IED Attack	IED Attack Against Coalition Forces
Government	Attack against Government/Political/Awakening Council
Indiscriminate	Indiscriminate Attack
Iraqi Security Forces	Attack against Iraqi Military Forces
Police	Attack against Iraqi Police
Religious Leader	Attack against Religious Leader (Sunni or Shia)
Tribal Leader	Attack against Tribal Leader (Sunni or Shia)
Unknown	Attack details do not permit classification

Table 2. Event Categories.

Based upon the contents of the Target field, an event from the predetermined categories is manually assigned to each record. The resulting table now contains the three fields required by the model (Date, Region (ID Field), and Event). This table along with the IED casualty table described previously can now be imported into the data mining tool for further manipulation and subsequently used as input to the sequence detection model.

Further data manipulation was carried out using the data preparation modules of the data mining tool used for this effort, SPSS Clementine. These manipulations included filtering out unused attributes, merging the IED casualty data with the Iraqi casualty data, removing records with null values, removing records with "Unknown" values in the event field, removing duplicate records, removing records of provinces with no IED attacks,

transforming the date field to a numeric Julian date, and sorting the data to improve the performance of the model.

b. Time Series

The goal of the time series model is to forecast the number of attacks in the immediate future months and identify what predictor variables (if any) influence the number of attacks, and therefore could be used by the model to increase the accuracy of its prediction.

Three sources of data were used for the time series model. The data needed for the model include past number of monthly attacks as well as data on any recurring events such as Ramadan and predictor variables such as coalition force levels that may have an effect on the number of the attacks.

The number of attacks that occurred each month was not directly available from the unclassified sources used for this research, but had to be derived from data obtained from icasualties.org.

(1) IED Casualty Data. As indicated, data obtained from icasualties.org consisted of deaths caused by IED and is presented in tabular format with columns of included Date, Name, Place of Death - Province, and Cause of Death. The individual names are hyperlinked to the U.S. Department of Defense Public Affairs website that includes a press release describing the details of the attack. A screen capture of data for the month of September 2003 is presented as Figure 11.

Sep-03

Date	Name	Place of Death - Province	Cause of Death
29-Sep-2003	Staff Sergeant Christopher E. Cutchall	Habbaniyah - Anbar	Hostile - hostile fire - IED attack
20-Sep-2003	Staff Sergeant Frederick L. Miller Jr.	Ramadi - Anbar	Hostile - hostile fire - IED attack
14-Sep-2003	Sergeant Trevor A. Blumberg	Fallujah - Anbar	Hostile - hostile fire - IED attack
09-Sep-2003	Specialist Ryan G. Carlock	Baghdad (NE of)	Hostile - hostile fire - IED attack
01-Sep-2003	Staff Sergeant Joseph Camara	Baghdad (south of)	Hostile - hostile fire - IED attack
01-Sep-2003	Sergeant Charles Todd Caldwell	Baghdad (south of)	Hostile - hostile fire - IED attack

Figure 11. Deaths Caused by IEDs - September 2003

The data represented by Figure 11 was imported into an Excel table for the initial data preparation efforts. The data needed from this table is the month/year and the number of attacks that occurred. The data shows there were six fatalities caused by IED attacks in September 2003, but an analysis of the Department of Defense press releases obtained by clicking on each of their names indicate that Staff Sergeant Camara and Sergeant Caldwell were killed in the same attack. So for the month of September 2003, there were five attacks. This manual process was repeated for each month from July 2003 to November 2008 to obtain the number of IED attacks that resulted in at least one fatality.

(2) Religious Holiday. The religious holiday that was chosen for this research as a potential predictor event was that of the Islamic holy month of Ramadan. Ramadan is a lunar based religious holiday that starts on different days of the calendar year and lasts approximately 30 days. The time series model for this research is aggregated to the month, so the start month and ending month of Ramadan is the data that was collected for

the years 2003 through 2008. If any days of Ramadan occurred during a calendar month, the value for the field in the table is set to one, if no days of Ramadan occurred during a given month, the value is set to zero. In 2003, Ramadan started in the month of October and ended in November. Therefore, the value for this field for these two months was set to one. For September 2003, the value of the field was set to zero.

(3) Coalition Force Strength. Coalition force strength data was used as a potential predictor variable to improve the accuracy of forecasting the number of IED attacks conducted by insurgents. Exact coalition force strength data is classified. However, unclassified numbers have been collected by the Saban Center for Middle East Policy and published as part of their Iraq Index. Specifically, the report contains data on the number of U.S. forces in Iraq, the number of international troops in Iraq, and the combined total of coalition troop levels in Iraq. The data was imported into an Excel spreadsheet for initial manipulation. For the month of September 2003, the number of coalition forces in Iraq was 156,000.

The time series model input record for the month of September 2003 is included as Table 3.

Month	IED Events	Ramadan	Coalition Strength
Sep-03	5	0	156,000

Table 3. September 2003 Time Series Input Data.

The output of the data preparation phase is the final data set (data that will be fed into the modeling tool) from the initial raw data collected from the identified data sources.

The next chapter focuses on applying modeling techniques to the data sets generated by the data preparation phase, and adjusting the parameters of the models to yield meaningful results. The chapter will also briefly discuss the evaluation and deployment of the generated models.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. MODELING, EVALUATION, AND DEPLOYMENT

A. INTRODUCTION

The previous chapter discussed the application of the first through third steps of the CRISP-DM methodology, namely business understanding, data understanding, and data preparation to the IED problem.

This chapter addresses the application of the remaining three steps of the methodology with emphasis on the fourth and vital step of the methodology, namely modeling.

This chapter is organized as follows: Section B discusses the application of the modeling step. It includes a discussion of the sequential pattern detection model followed by a discussion of the time series model. Sections C and D discuss the evaluation and deployment phases of the CRISP-DM methodology respectively.

B. MODELING

The purpose of the modeling phase of the CRISP-DM methodology is to select, build, and assess models that support the operational objective. For this research, sequential pattern detection and time series are the two models that have been chosen and are discussed in this section.

1. Sequential Pattern Detection

a. Model Selection

A sequence is a list of item sets that tends to occur in a predictable order. For example, the occurrence

of certain religious events and political events are almost always followed by some sort of insurgency attack. Sequence modeling detects frequent sequences and generates a model that can be used to make predictions. Sequence modeling used in this research is based on the Continuous Association Rule Mining Algorithm (CARMA) association rules algorithm, which uses an efficient two-pass method for finding sequences. The results of the research will be in the form of a model that identifies commonly occurring or high confidence sequences. These sequences are presented in the following format:

Consequent ⇐ Antecedent 1 Antecedent 2 … Antecedent N

For example:

Consequent	Antecedent 1	Antecedent 2
Coalition IED Attack	Religious Holiday	Iraqi Police Attack

This sequence tells us that if there is a religious holiday, and later an Iraqi police attack, it is likely to have an IED attack against coalition forces. In addition to generating sequences, the analysis provides evaluation measures of the support and confidence of the generated sequences.

b. Testing Model Design

The sequence detection model for this research is based upon the CARMA Algorithm proposed by Christian Hidber in 1999 (Hidber, 1999). CARMA outperforms previous association rule algorithms (Apriori and DIC) in certain situations and uses less computer memory. The mechanics of

how CARMA works is beyond the scope of this thesis, but CARMA is an academically well-accepted implementation for finding rule associations in large data sets.

c. Build Model

Building the model involves determining model parameters and optimizing the model for performance or results. The parameters that can be set in the SPSS sequence modeling node used to implement the algorithm include:

(1) Minimum rule support. Rule support refers to the proportion of cases that contain the entire sequence. A higher minimum rule support identifies more commonly occurring sequences.

(2) Minimum rule confidence. Rule confidence refers to the proportion of cases with the antecedents that also include that specific consequent. Increasing the minimum rule confidence would reduce the number of sequences and eliminate uninteresting ones.

(3) Maxduration constraint. The maxduration constraint specifies the maximum allowed time difference between the first and last item set.

(4) Maximum sequence size. Maximum sequence size is the maximum number of distinct item sets, as opposed to items, in a sequence.

(5) The maxspan constraint. Maxspan constraint specifies the maximum allowed time difference between the latest and earliest occurrences of events in a sequence.

(6) The mingap and maxgap constraints. These constraints restrict the time difference between two

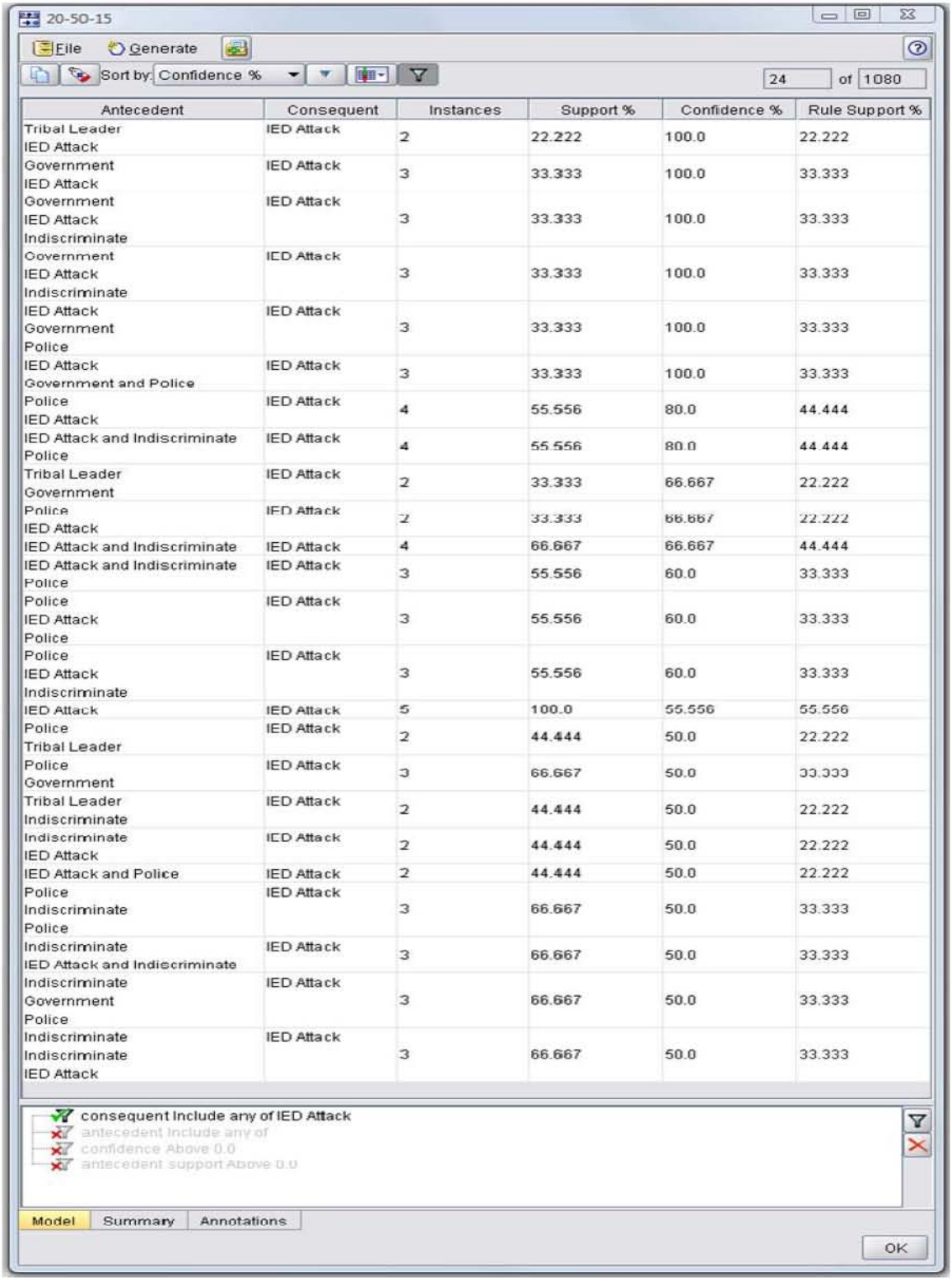
consecutive elements of a sequence. The maxgap constraint specifies the maximum amount of time that separate item sets in a sequence. The mingap constraint specifies the minimum amount of time that separate item sets in a sequence.

d. Assess Model

Multiple models were created by varying four of the key parameters (minimum rule support, minimum rule confidence, maxduration constraint, and the maxgap constraint). Maximum sequence size was set to ten for all models. Mingap and maxspan were not constrained.

Minimum rule support was set at either twenty percent, twenty-five percent, or thirty percent. Minimum rule confidence was set at either fifty percent, sixty percent, or seventy percent. The maxduration constraint was set at either five, ten, or fifteen days, and the maxgap constraint was either not constrained or set to five days. A discussion of the models follows.

(1) Sequential Model 20-50-15. This is the least restrictive model generated and is the model with the settings of minimum rule support of twenty percent, minimum rule confidence of fifty percent, and maxduration constraint of fifteen days. The results of the model are shown in Figure 12.



Antecedent	Consequent	Instances	Support %	Confidence %	Rule Support %
Tribal Leader IED Attack	IED Attack	2	22.222	100.0	22.222
Government IED Attack	IED Attack	3	33.333	100.0	33.333
Government IED Attack Indiscriminate	IED Attack	3	33.333	100.0	33.333
Government IED Attack Indiscriminate	IED Attack	3	33.333	100.0	33.333
IED Attack Government Police	IED Attack	3	33.333	100.0	33.333
IED Attack Government and Police	IED Attack	3	33.333	100.0	33.333
Police IED Attack	IED Attack	4	55.556	80.0	44.444
IED Attack and Indiscriminate Police	IED Attack	4	55.556	80.0	44.444
Tribal Leader Government	IED Attack	2	33.333	66.667	22.222
Police IED Attack	IED Attack	2	33.333	66.667	22.222
IED Attack and Indiscriminate	IED Attack	4	66.667	66.667	44.444
IED Attack and Indiscriminate Police	IED Attack	3	55.556	60.0	33.333
Police IED Attack Police	IED Attack	3	55.556	60.0	33.333
Police IED Attack Indiscriminate	IED Attack	3	55.556	60.0	33.333
IED Attack	IED Attack	5	100.0	55.556	55.556
Police Tribal Leader	IED Attack	2	44.444	50.0	22.222
Police Government	IED Attack	3	66.667	50.0	33.333
Tribal Leader Indiscriminate	IED Attack	2	44.444	50.0	22.222
Indiscriminate IED Attack	IED Attack	2	44.444	50.0	22.222
IED Attack and Police	IED Attack	2	44.444	50.0	22.222
Police Indiscriminate Police	IED Attack	3	66.667	50.0	33.333
Indiscriminate IED Attack and Indiscriminate	IED Attack	3	66.667	50.0	33.333
Indiscriminate Government Police	IED Attack	3	66.667	50.0	33.333
Indiscriminate Indiscriminate IED Attack	IED Attack	3	66.667	50.0	33.333

Figure 12. Sequence Model 20-50-15.

There were 1,080 sequences identified by the model, however, only twenty-four of those have a consequent of IED attack against coalition forces. The sequence detection investigates all time ordered associations among the analysis variables and not only those related to a specific outcome category. However, after the sequences are generated, filters can be applied to show only those outcome categories of interest. The sequences that do not have a consequent of an IED attack have been filtered out.

The first rule indicates that when an attack against a tribal leader occurs and later an IED attack occurs, then there is a one hundred percent likelihood that an IED attack will occur within the maxduration constraint of fifteen days. The second rule listed shows that when there is an attack against a government target, followed by an IED attack, there is a one hundred percent likelihood that an IED attack will occur within fifteen days of the first event in the events set.

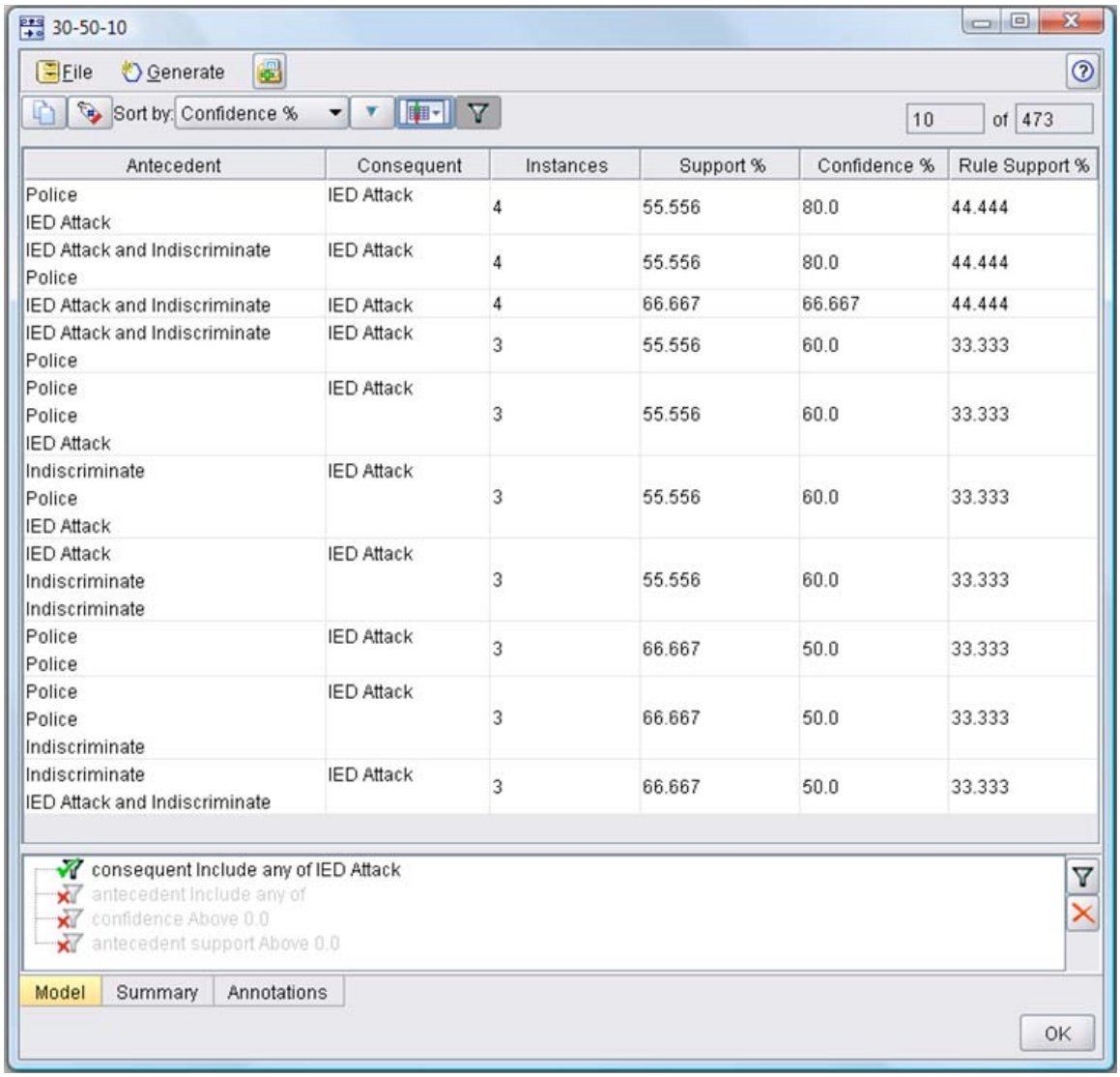
Unfortunately, the sequences identified by this model are not particularly useful, given the maxduration constraint of fifteen days. Given enough time, there is sufficient violence in Iraq for any type of attack to be an antecedent for any other type of attack.

(2) Sequential Model 30-70-5. This is the most restrictive model generated to see how that compares with the first model. The settings for the most restrictive model are minimum rule support of thirty percent, minimum rule confidence of seventy percent, and maxduration constraint of five days. This model identified sixty sequences; however, none of them have an IED attack

as a consequent. Neither of the first two models discussed have much use to the operational commander in the field as they are either too generic or too restrictive.

(3) Sequential Models with Maxduration of 5. There were nine models with a maxduration constraint of five days and only one sequence association rule was generated in those nine models. That rule indicates when there is an attack against a tribal leader, followed by an indiscriminate attack, there will an IED attack with a sixty-six percent likelihood within five days.

(4) Sequential Model 30-50-10. Since a five-day maxduration constraint is too short to generate any rules and a fifteen day maxduration constraint generates too many rules, the models with a maxduration constraint of ten days will be examined next. The model results using the settings of minimum rule support of thirty percent, minimum rule confidence of fifty percent, and maxduration constraint of ten days are shown in Figure 13.



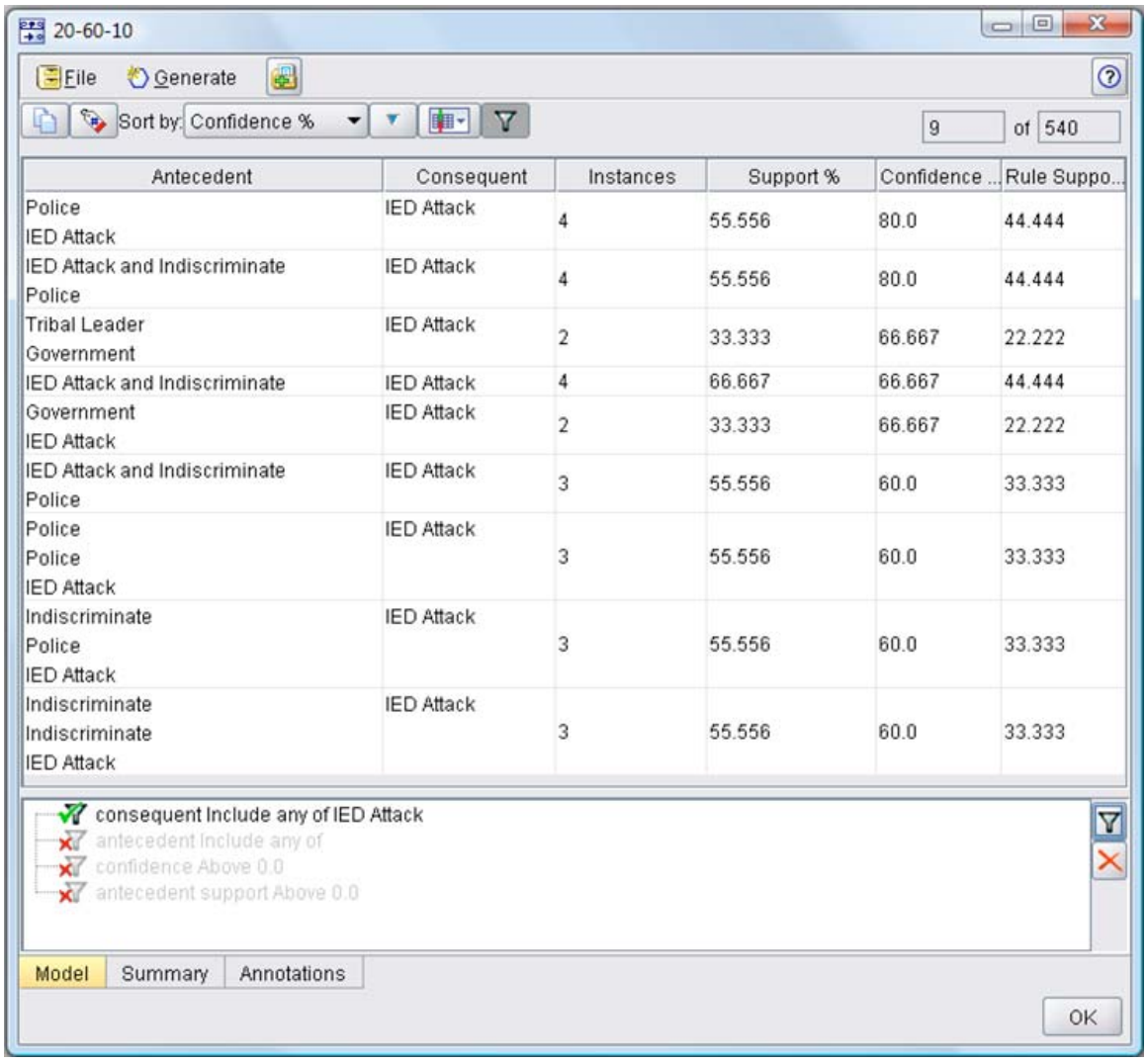
Antecedent	Consequent	Instances	Support %	Confidence %	Rule Support %
Police IED Attack	IED Attack	4	55.556	80.0	44.444
IED Attack and Indiscriminate Police	IED Attack	4	55.556	80.0	44.444
IED Attack and Indiscriminate	IED Attack	4	66.667	66.667	44.444
IED Attack and Indiscriminate Police	IED Attack	3	55.556	60.0	33.333
Police Police IED Attack	IED Attack	3	55.556	60.0	33.333
Indiscriminate Police IED Attack	IED Attack	3	55.556	60.0	33.333
IED Attack Indiscriminate Indiscriminate	IED Attack	3	55.556	60.0	33.333
Police Police	IED Attack	3	66.667	50.0	33.333
Police Police Indiscriminate	IED Attack	3	66.667	50.0	33.333
Indiscriminate IED Attack and Indiscriminate	IED Attack	3	66.667	50.0	33.333

Figure 13. Sequence Model 30-50-10.

The model identified 473 sequences with ten of them having an IED attack against coalition forces as the consequent. The first rule indicates that when an attack against Iraqi police occurs followed by an IED attack, then it is eighty percent likely that an IED attack will occur within the maxduration constraint of ten days. The second rule shows that when there is an IED attack against coalition forces and an indiscriminate attack on

the same day, followed by a police attack then there is an eighty percent likelihood of an IED attack within the maxduration constraint of ten days.

(5) Sequential Model 20-60-10. This model was created using the settings of minimum rule support of twenty percent, minimum rule confidence of sixty percent, and maxduration constraint of ten days. The resulting sequence association rules are shown in Figure 14.



Antecedent	Consequent	Instances	Support %	Confidence ...	Rule Suppo...
Police IED Attack	IED Attack	4	55.556	80.0	44.444
IED Attack and Indiscriminate Police	IED Attack	4	55.556	80.0	44.444
Tribal Leader Government	IED Attack	2	33.333	66.667	22.222
IED Attack and Indiscriminate	IED Attack	4	66.667	66.667	44.444
Government IED Attack	IED Attack	2	33.333	66.667	22.222
IED Attack and Indiscriminate Police	IED Attack	3	55.556	60.0	33.333
Police Police IED Attack	IED Attack	3	55.556	60.0	33.333
Indiscriminate Police IED Attack	IED Attack	3	55.556	60.0	33.333
Indiscriminate Indiscriminate IED Attack	IED Attack	3	55.556	60.0	33.333

Figure 14. Sequence Model 20-60-10.

The model identified 540 sequences with nine of them having an IED attack against coalition forces as the consequent. The first two rules are the same as those generated by the previously discussed model. The third rule indicates that when an attack against a tribal leader is followed by an Iraqi government attack, then it is sixty-six percent likely that an IED attack will occur within the maxduration constraint of ten days.

The models discussed so far have not been constrained by the maxgap parameter. As defined earlier, the maxgap constraint specifies the maximum amount of time that separate item sets in a sequence. A model with the maxgap constraint set to five days was generated and is discussed next.

(6) Sequential Model 20-50-15-5Gap. This model is similar to Sequential Model 20-50-15 with a minimum rule support of twenty percent, minimum rule confidence of fifty percent, and maxduration constraint of fifteen days. The maxgap constraint of five days was added to the model.

There were 345 sequences identified by the model, however, only twenty-three of those have a consequent of IED attack against coalition forces. The first rule indicates that when an IED attack occurs followed by three subsequent indiscriminate attacks within five days of each other, then there is a one hundred percent likelihood that an IED attack will occur within five days of the third indiscriminate attack in the sequence. The second rule generated shows that when there is an IED attack and an indiscriminate attack on the same

day followed by two subsequent indiscriminate attacks, then there is a one hundred percent likelihood of an IED attack occurring within five days of the last indiscriminate attack in the sequence.

Even though some of the sequences identified by this model are different than the sequences identified by the non-maxgap constrained model, neither model is more "correct" than the other model, they both identify sequences that have an IED attack as a consequent.

The models generated provide an analysis of the sequences of events that have occurred and may provide some useful intelligence to the operator in the field about whether an IED attack is imminent given a sequence of events that have happened already. Subject matter experts in the field are needed to help with the validation of the rules that the models have indentified to further improve the models.

2. Time Series

a. Model Selection

Time series analysis examines historical data to determine if there are any patterns that can be used to predict where future values of the data series are likely to fall. For example, by using historical sales to forecast future sales for automobiles, a car manufacturer would be able to decide where best to allocate production resources.

The two most important methodologies for conducting time series analysis are exponential smoothing and autoregressive integrated moving average (ARIMA).

Exponential smoothing uses weighted historical values to forecast future values. A simple moving average model would assign an equal weight to each of the previous data points when predicting the next value. Exponential smoothing assigns an exponentially decreasing weight to older values, giving the more recent data a larger influence on the model. Exponential smoothing is pure time series model because the only independent variable allowed is time.

The ARIMA methodology is a more sophisticated way of modeling time series data because it allows the inclusion of one or more independent variables, called predictors, that could improve the ability of the model to forecast the future values of the series. For example, one of the predictors that this research examines is the coalition force strength in Iraq and what effect, if any, the number of troops has on the number of fatal IED attacks that occur each month. Additionally, ARIMA models can incorporate events and interventions. An event is a type of predictor that occurs on a predicable basis. Ramadan is an event that will be used in this research. An intervention is another type of predictor variable that is used for single incidents that have occurred in the past. No intervention variables are used in the subsequent time series models.

b. Testing Model Design

Exponential smoothing and ARIMA are two standard statistical models from a design perspective. The models themselves have been tested extensively and have shown to be valid.

In addition to the above models, the data mining tool used for this research, SPSS Clementine, has a third modeling option, called Expert Modeler, which automatically finds the best-fitting model for each target variable to be forecasted, based on a number of goodness-of-fit measures.

c. Build Model

SPSS Clementine implements multiple variants of exponential smoothing:

- Simple - This model is appropriate for series in which there is no trend or seasonality.
- Holts linear trend - This model is appropriate for series in which there is a linear trend and no seasonality.
- Browns linear trend - This model is appropriate for series in which there is a linear trend and no seasonality. Its relevant smoothing parameters are level and trend, but, in this model, they are assumed to be equal. Brown's model is therefore a special case of Holt's model.
- Damped trend - This model is appropriate for series with a linear trend that is dying out and with no seasonality.
- Simple seasonal - This model is appropriate for series with no trend and with a seasonal effect that is constant over time.
- Winters' additive - This model is appropriate for series with a linear trend and a seasonal effect that is constant over time.
- Winters' multiplicative - This model is appropriate for series with a linear trend and a seasonal effect that changes with the magnitude of the series.

The ARIMA settings available to the analyst are autoregressive (p), difference (d), and moving average (q). For this research, the expert modeler function was chosen,

which automatically identifies and estimates the best-fitting ARIMA or exponential smoothing model, thus eliminating the need to identify an appropriate model through trial and error. The output of the model is a forecasted value of the target variable together with a series of statistical goodness-of-fit measures. Some of the statistical measures provided are:

(1) Stationary R^2. Stationary R^2 is a goodness of fit statistic and can range from negative one to positive one. It provides an estimate of the proportion of the total variation in the series that is explained by the model. The closer the value is to positive one the better the model fits the data.

(2) R^2. R^2 is a goodness of fit statistic and can range from negative one to positive one. It is an estimation of the total variation in the time series that can be explained by the model. The closer the value is to positive one the better the model fits the data.

(3) Root Mean Square Error (RMSE). RMSE is a measure of the difference between the actual data points and the predicted points that the model generated and is expressed in the same units as those used for the series itself. A smaller difference indicates a better model.

(4) Mean Absolute Percentage Error (MAPE). MAPE is a measure of the average percentage error between actual data points and predicted data points. A smaller value indicates a better model.

(5) Mean Absolute Error (MAE). MAE is a measure of the average difference between actual values and

predicted values expressed in the same units as those used for the series itself. A smaller value indicates a better model.

(6) Maximum Absolute Percentage Error (MaxAPE). MaxAPE is the maximum percentage error between all predicted values compared and their corresponding actual values. A smaller value indicates a better model.

(7) Maximum Absolute Error (MaxAE). MaxAE is the largest error between all the predicted values compared and their corresponding actual values expressed in the same units as those used for the series itself. A smaller value indicates a better model.

(8) Normalized Bayesian Information Criterion. A general measure of the overall fit of a model that attempts to account for model complexity.

(9) Autocorrelation Function (ACF). This measures the correlation between series values that are k units apart, where k is set by the analyst. ACF values range from positive one to negative one.

(10) Partial Autocorrelation Function (PACF). This measures the correlation between series values that are k units apart, and takes into consideration the values of the intervening intervals. PACF values range from positive one to negative one.

d. Assess Model

Four time series models where created to forecast the number of IED attacks. The initial model did not consider any predictor variables or events. Subsequent

models considered the level of coalition force strength as a predictor variable as well as the religious holiday of Ramadan as an event.

(1) Time Series without Predictors. The first model built used the expert modeler without the addition of any predictor variables. The outputs of the model are forecasted values of the number of IED attacks, which can be plotted on a graph with the actual values of the IED attacks in the estimation period, together with the statistical goodness-of-fit measures described in the previous section. The graph for the basic model without predictor variables showing the forecasted values and the actual values is shown in Figure 15.

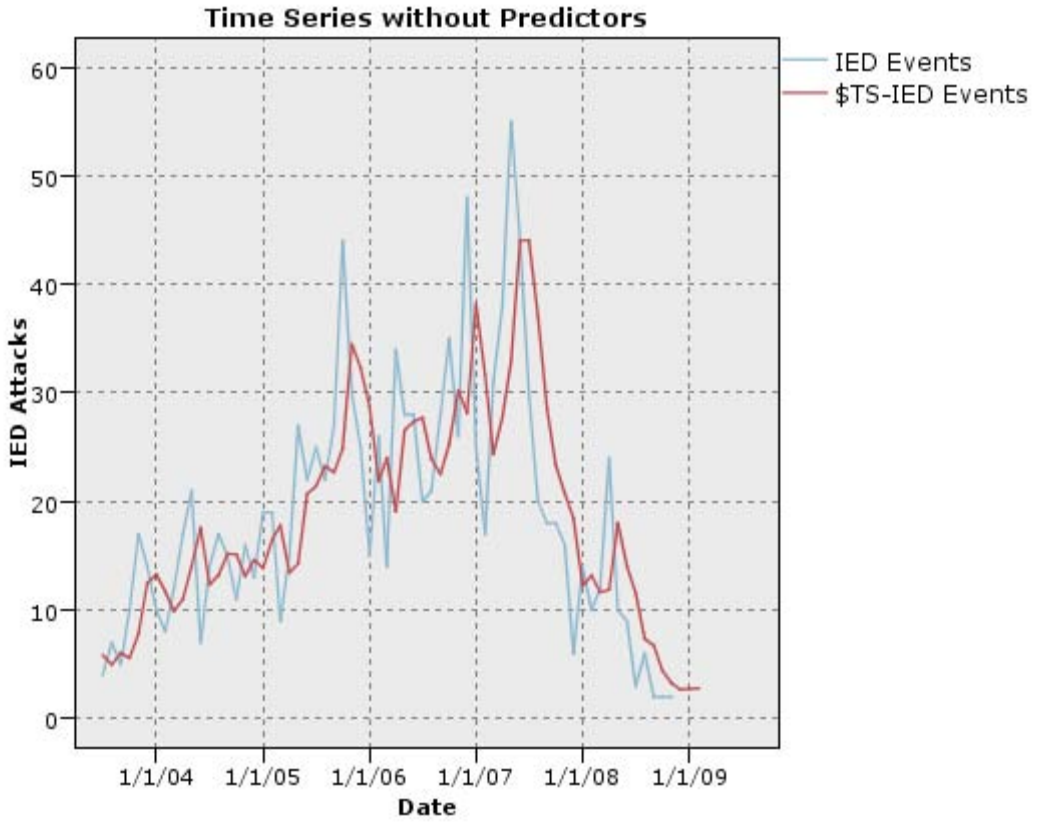


Figure 15. Time Series Model without Predictors.

The blue line represents the number of actual IED attacks that occurred and the red line represents the number of IED attacks predicted by the time series model. The expert modeler feature chose the simple seasonal algorithm of the exponential smoothing model as the best fitting model, based on the goodness-of-fit measures shown in Table 4.

Model	Simple Seasonal
Predictors	0
Stationary R^2	0.659
R^2	0.598
RMSE	7.344
MAPE	42.322
MAE	5.822
MaxAPE	280.276
MaxAE	16.842

Table 4. Time Series Model without Predictor Variables.

The results of the model are not that good. The Mean Absolute Percentage Error (MAPE), which indicates how much a dependent series varies from its model-predicted level, has a value of forty-two percent. The Mean Absolute Error (MAE), which measures how much the series varies from its model-predicted level in the original series units shows an average error of almost six IED attacks in a given

month. The R^2 value of 0.598 also does not indicate a good fit, a value of 0.8 or higher would be preferred. If this model is not that good of a prediction for the number of attacks, when the predictor variables are added in, does the model improve?

(2) Time Series with Coalition Force Predictor. The second model to be analyzed is a model that adds the predictor variable of coalition force strength to see whether the number of boots on the ground improves the ability of the model to forecast the number of IED attacks that occur in each month. The output of this model is presented in Figure 16.

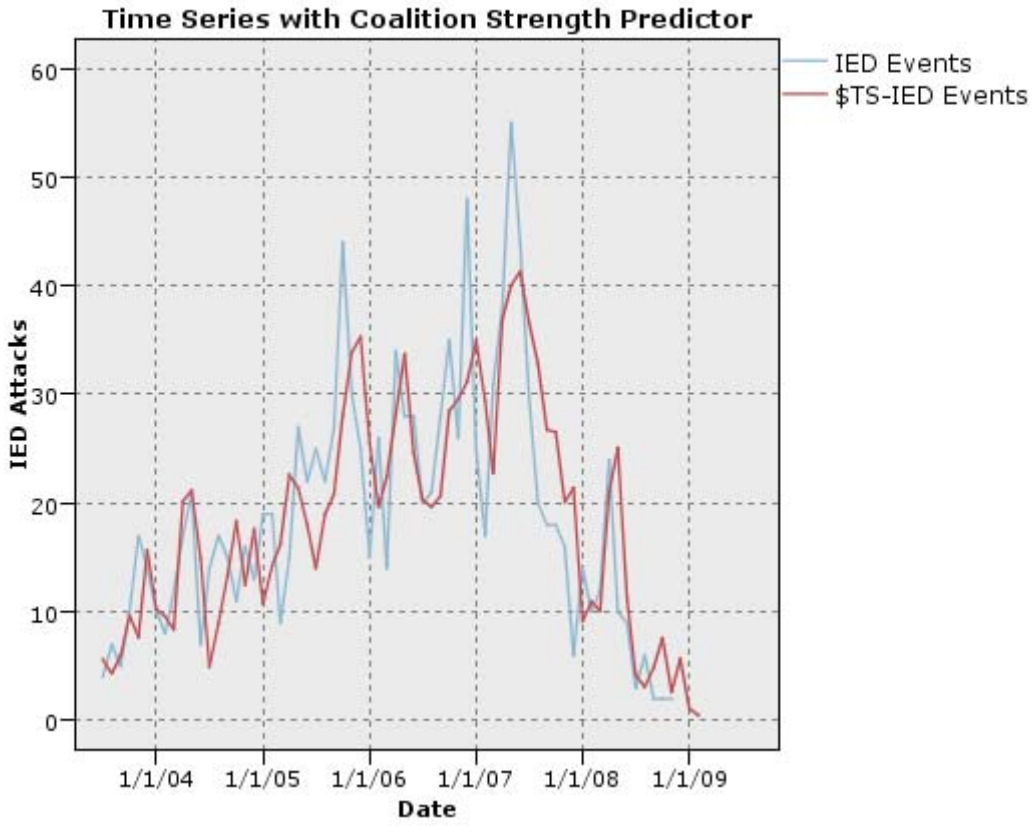


Figure 16. Time Series Model with Coalition Forces Predictor.

The blue line represents the number of actual IED attacks that occurred and the red line represents the number of IED attacks that the predicted values. The output of this model looks identical to the previous model without a predictor variable. An examination of the resulting statistical measures of the model verify the rejection of the predictor variable by the expert modeler. The expert modeler again chose the simple seasonal algorithm of the exponential smoothing model as the best fitting model for the data. The statistical measures for the model are shown in Table 5.

Model	Simple Seasonal
Predictors	0
Stationary R^2	0.659
R^2	0.598
RMSE	7.344
MAPE	42.322
MAE	5.822
MaxAPE	280.276
MaxAE	16.842

Table 5. Time Series Model with Coalition Forces Predictor.

The statistical measures confirm that adding the predictor variable of coalition force strength does not result in a better predictor model, because the expert

modeler rejected it and chose the model without any predictors based upon the statistics of both models.

(3) Time Series with Ramadan Event. The third model created added the religious holiday Ramadan as an event and removed the coalition forces predictor. The expert modeler also rejected the third model in favor of the simple seasonal model.

(4) Time Series with both Coalition Force Predictor and Ramadan Event. The fourth and final model combined both predictor variables of coalition force strength and Ramadan. The results were again the same as the expert modeler chose the simple seasonal model with no predictors over a model with two predictor variables. This leads to the conclusion that Ramadan and coalition force strength are not a significant influence on the ability of the model to forecast the number of IED attacks that occur in a given month.

The expert modeler option is a valuable time saving feature that computes multiple models simultaneously and then selects the best model based on goodness-of-fit measures. It would be prudent to at least run a few models in manual mode to verify that the expert modeler is indeed selecting the best model.

(5) Holts Linear Trend Model. This model has no predictor variables and instead of allowing the expert modeler to choose the best fit model, the Holts linear trend model is manually selected. The resulting model is shown as Figure 17.

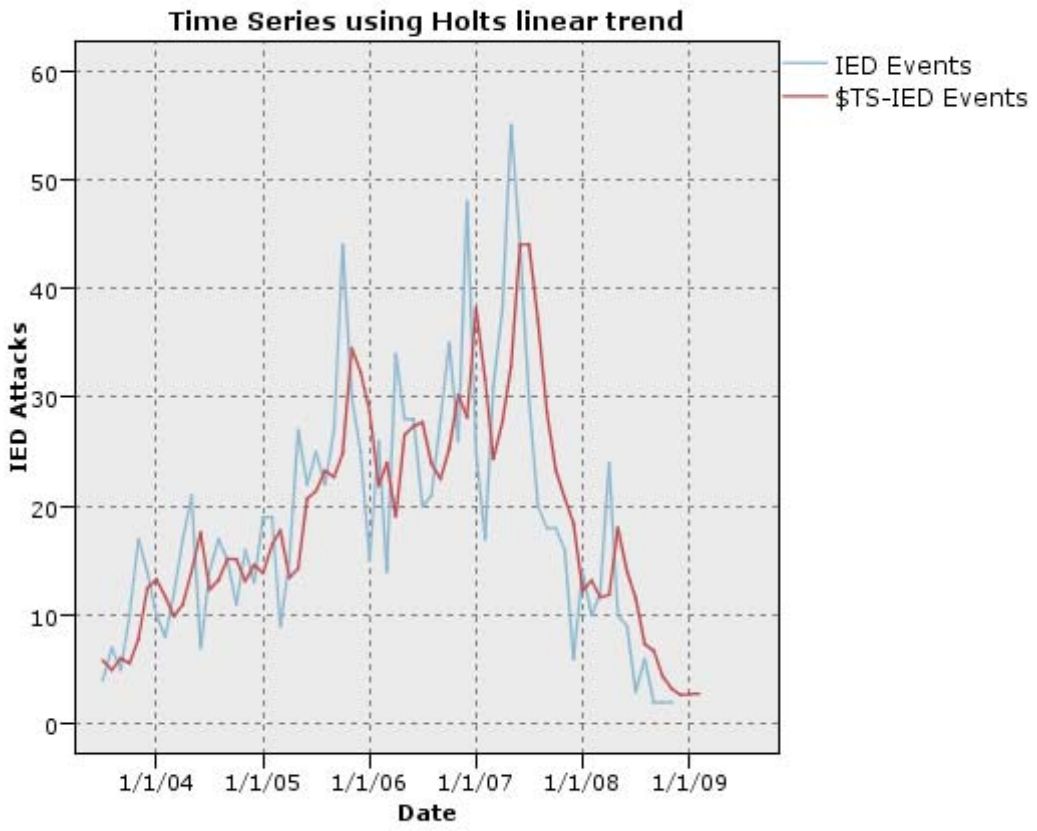


Figure 17. Time Series Model using Holts linear trend.

The resulting graph is visually indistinguishable from the simple linear trend model that the expert modeler chose. The statistics bear out the slight improvement the simple seasonal model offers over the Holts linear trend model. A side-by-side comparison of the statistics is presented in Table 6.

Model	Simple Seasonal	Holts linear trend
Predictors	0	0
Stationary R^2	0.659	0.656
R^2	0.598	0.473
RMSE	7.344	8.406
MAPE	42.322	45.927
MAE	5.822	6.361
MaxAPE	280.276	286.155
MaxAE	16.842	22.07

Table 6. Time Series Model Statistics (Simple vs. Holts).

(6) Coalition Forces ARIMA Model. As an additional verification of the expert modeler's rejection of more advanced ARIMA models, the next model presented is an ARIMA model using the coalition force predicator variable and setting the ARIMA model criteria of autoregressive, difference, and moving average all set to zero. The resulting predicted values and actual values are shown in Figure 18.

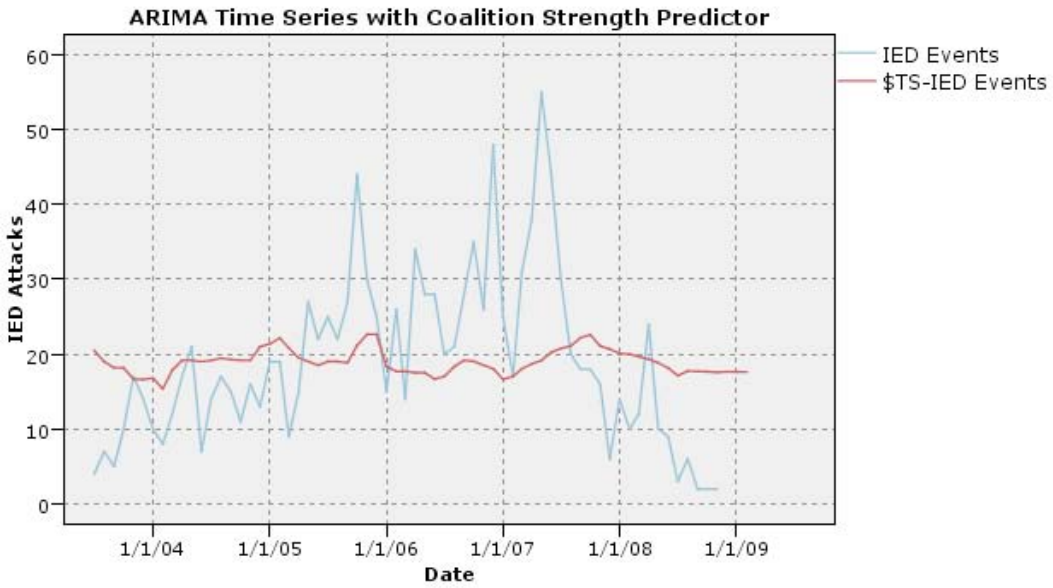


Figure 18. ARIMA Time Series with Coalition Strength Predictor.

It is immediately obvious from the resulting graph that the manual ARIMA model is not a good fit to the observed data. The statistical results are presented in Table 7 for comparison to the simple seasonal model that the expert modeler chose.

Model	Simple Seasonal	ARIMA (0,0,0)
Predictors	0	1
Stationary R^2	0.659	0.02
R^2	0.598	0.02
RMSE	7.344	11.457
MAPE	42.322	98.806
MAE	5.822	8.97
MaxAPE	280.276	785.699
MaxAE	16.842	35.843

Table 7. Time Series Model Statistics (Simple vs. ARIMA).

The simple seasonal model chosen by the expert modeler has better values for the statistical measures in each category than the ARIMA model verifying that the expert modeler indeed selected the best fitting model, based on the statistical measures.

C. EVALUATION

The purpose of the evaluation phase is to compare the data mining results to the original goals of the data mining project set forth in the business understanding phase to determine if the success criteria have been met.

For this project, sequential pattern detection modeling discovered sequences that have an IED attack

against coalition forces as a consequent of the sequence. It is difficult to assess the usefulness of the discovered sequences without subject expert opinion or being deployed to the operational theater. Subject matter experts in the field are needed to provide feedback on the validity of the sequences that have been discovered. The models show that this line of research is valid and can be an additional tool to assist deployed forces with countering the IED threat.

D. DEPLOYMENT

The deployment phase is the final phase of the CRISP-DM methodology and consists of the deployment, maintenance and upkeep, and final reporting of the data mining project. There are no plans to deploy the models generated by this research, but recommendations will be made in the following chapter regarding follow on work.

This chapter presented the last three steps of the CRISP-DM methodology with particular emphasis on the design, implementation, and results of modeling. The next chapter summarizes the thesis, and discusses conclusions, lessons learned, and recommended areas for further research.

V. SUMMARY, CONCLUSIONS, LESSONS LEARNED, AND FUTURE WORK

A. SUMMARY

This thesis implemented industry standard data mining techniques to support the business (operational) objective of reducing the number of IED attacks that cause casualties to coalition forces in Iraq. The two data mining techniques used in this thesis were sequence pattern detection and time series analysis.

Sequence pattern detection seeks to identify potential sequences of events that could be used to predict when IED attacks will occur against coalition forces. Time series analysis examined if there were trends in the number of IED attacks that occur monthly, and if predictor variables could be used to improve the prediction model.

Chapter II of the thesis discussed the common types of data mining tasks, the data mining models and algorithms to support those tasks, and described an industry-wide standardized process approach, called CRISP-DM, to data mining.

The third chapter addressed the implementation of the first, second, and third steps of the CRISP-DM methodology, namely business understanding, data understanding, and data preparation to the IED problem. Discussion topics included data categories and sources, file format for data mining, and data preparation/manipulation.

The fourth chapter presented the last three steps of the CRISP-DM methodology with particular emphasis on the design, implementation, and results of modeling.

B. CONCLUSIONS

The objective of this research was to develop a predictive model for the timing and frequency of IED attacks using sequence pattern detection and time series analysis. The following research questions were proposed:

1. Can data mining techniques/approaches be useful in predicting the timing, frequency, and number of IED attacks?

2. Can the models be deployed to operational commands and be used by deployed personnel?

3. How accurate can these models be? Models will be evaluated for confidence, support, and goodness of fit.

As a result of the research conducted by this thesis, we can make the following conclusions. The answer to the first research question is a definite yes, data mining techniques can be useful as an additional tool to the commander in the field to mitigate the effects of IED attacks (defeat the device).

The answer to the second question is that although the models in their current state are not deployable to operational commands, they eventually could be deployed following a rigorous process of evaluation, validation, verification, and selection of relevant and appropriate rules. SPSS Clementine does support creating stand alone executable modules for deployment of the rules in the theater of operations.

The answer to the third question is that while some of the sequence pattern models appear to be good and useful, subject matter experts are needed to review the results and to suggest improvement for future modeling efforts. No prediction model can be one hundred percent accurate, but useful information can be drawn from the models and as the model iteratively gets better, the value of the knowledge gained will increase.

The time series models do not exhibit a good fit to the actual data. While is it certainly possible that coalition force strength and Ramadan, as well as other predictor factors, may have an influence of the number of monthly IED attacks, the data used in this thesis and the generated results did not support the idea that coalition forces or Ramadan were a major factor in predicting the number of monthly attacks. More research is needed to identify potential factors that influence the predictability of the number of monthly IED attacks.

C. LESSONS LEARNED

There were five main lessons learned during this thesis:

- **Data preparation for data mining is time consuming**

Data acquisition, preparation, and understanding is a huge undertaking. Even though all of the data was from the unclassified domain, it took time to find the quantity and quality of data that needed to generate the models. Once the data was located, a significant amount of time was required to format the data into a useable form for the

data mining tasks. A rough estimate of the data acquisition and preparation effort is sixty to seventy percent of the total effort.

- **Source data classification does not have to be a deterrent to good research**

Information available in the unclassified domain is sketchy and not as relevant and accurate as information that is in the classified domain. However, the goal of this thesis was to provide a theoretical framework of how sequence detection models and time series analysis could be used to defeat IEDs, regardless of the relevance and accuracy of the source data.

- **Using a methodology is crucial in guiding any data mining effort**

Following the CRISP-DM methodology facilitated a thorough and orderly progression through the data mining project. Following an iteratively structured approach for any future data mining efforts is vital.

- **Adopting data mining tools that support all phases of the data mining methodology makes the data mining process more efficient**

Utilizing the features of the data mining tool for all phases of the data mining methodology makes the data mining process more efficient than using individual tools for individual tasks. Excel was used in the beginning of the project to manipulate and prepare the data. Clementine has more powerful data manipulation tools built in that saved time compared to the manual editing of data that was required in Excel.

- **Input from stakeholders and subject matter experts is vital to the success of the data mining effort**

Stakeholders and subject matter experts are needed to participate in the development of future data mining efforts. They could have provided insights, suggestions, and anecdotal evidence based up actual field observations that are not available to academic researchers.

D. FUTURE WORK

This research suggests five areas for further research:

- **Validating resulting sequences by domain experts**

Resulting sequences need to be evaluated by domain experts to identify those sequences that are novel, useful, and actionable. Domain experts could also help improve the models by proving a fresh perspective on the problem and could be beneficial in the hypothesis generation phase for future modeling efforts.

- **Adding additional events as predictors to an IED attack**

Additional events need to be incorporated for both sequence pattern detection and time series models as predictors to an IED attack. These could include religious holidays, political events, as well as a breakdown of indiscriminate, religious leaders, and tribal leaders attacks by religious affiliation (e.g., Sunni vs. Shia). Gathering observations from deployed personnel is recommended for generating new predictor variables to examine.

- **Using a smaller unit of analysis**

The unit of analysis for this research was the provinces of Iraq and no distinction was made between events that happened inside different geopolitical and socioeconomic zones inside of a province. Future research should break down the unit of analysis into smaller areas that offer a more homogenous population based upon religious affiliation and socioeconomic similarities instead of provincial boundaries.

- **Adding geographical analysis of events**

The event is this thesis were studied from a temporal perspective, trying to predict future events by analyzing past events. Future work could look at using event positional data to search for trends that are not apparent using time based analysis methods.

- **Using crime scene analysis software tools**

An IED attack against coalition forces is very similar to a crime scene in terms of evidence, witnesses, and timing. An interesting idea for future work is to utilize commercial crime scene analysis software to search for patterns in the attacks similar to the way law enforcement agencies look for criminal modus operandi.

LIST OF REFERENCES

Berry, M.J.A. and G.S. Linoff. *Data Mining Techniques*, 2d ed., Wiley Publishing Inc., 2004.

CRISP-DM 1.0, 1st ed., v.1. CRISP-DM Consortium, 2000.

Hidber, C. "Online Association Rule Mining," paper presented at the ACM SIGMOD International Conference on Management of Data, Philadelphia, Pennsylvania, 1-3 June 1999.

iCasualties: Iraq Coalition Casualty Count. http://icasualties.org/Iraq/index.aspx, February 2009.

Iraq Body Count, http://www.iraqbodycount.org/, February 2009.

Counter IED Technology. https://www.jieddo.dod.mil/CIEDTECHNOLOGY/CIEDTECHHOME.ASPX, March 2009.

Islamic Holidays - Religion Facts. http://www.religionfacts.com/islam/holidays.htm

Two Crows. Introduction to Data Mining and Knowledge Discovery, 3rd ed., v.1, Two Crows Corporation, 1999.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Joint Improvised Explosive Device Defeat Organization
   Crystal City, Virginia

4. Dr. Magdi Kamel
   Naval Postgraduate School
   Monterey, California

5. Mr. Albert Barreto
   Naval Postgraduate School
   Monterey, California